PONTIFICIAE ACADEMIAE SCIENTIARVM SCRIPTA VARIA 121

Edited by
A. Battro, S. Dehaene, M. Sánchez Sorondo and W. Singer

Neurosciences and the Human Person: New Perspectives on Human Activities

THE PROCEEDINGS OF THE WORKING GROUP • 8-10 NOVEMBER 2012

VATICAN CITY 2013

Neurosciences and the Human Person: New Perspectives on Human Activities

Pontificiae Academiae Scientiarvm Scripta Varia 121

The Proceedings
of the Working Group on

Neurosciences and the Human Person: New Perspectives on Human Activities

8-10 November 2012

Edited by

Antonio M. Battro
Stanislas Dehaene
Marcelo Sánchez Sorondo
Wolf J. Singer

EX AEDIBVS ACADEMICIS
IN CIVITATE VATICANA • MMXIII

The Pontifical Academy of Sciences
Casina Pio IV, 00120 Vatican City
Tel: +39 0669883195 • Fax: +39 0669885218
Email: pas@pas.va • Website: www.pas.va

The opinions expressed with absolute freedom during the presentation of the papers of this meeting, although published by the Academy, represent only the points of view of the participants and not those of the Academy.

ISBN 978-88-7761-106-2

PONTIFICIA ACADEMIA SCIENTIARVM • VATICAN CITY

Man's greatness lies in his capacity to think of God. And this means being able to live a conscious and responsible relationship with Him. However, the relationship is between two realities. God – this is my thought and this is my experience, but how many, yesterday and today, share it! – is not an idea, even though very lofty, fruit of man's thought. God is reality with a capital "R". Jesus reveals it – and lives the relationship with Him – as a Father of goodness and infinite mercy. Hence, God doesn't depend on our thought. Moreover, even when the life of man on earth should finish – and for the Christian faith, in any case, this world as we know it is destined to fail – man won't stop existing and, in a way that we don't know, nor will the universe that was created with him. Scripture speaks of "new heavens and a new earth" and affirms that, in the end, in the where and when that is beyond us, but towards which, in faith, we tend with desire and expectation, God will be "all in all". … Jesus was sent by Abba "to preach good news to the poor, to proclaim release to captives, and recovering of sight to the blind, to set at liberty those who are oppressed, to proclaim the acceptable year of the Lord" (Luke 4:18-19).

Pope Francis, Letter to Eugenio Scalfari, *La Repubblica*, 11 September 2013.


J. MEHLER
W.A. NOWAK
S. ARBER
T.K. BEHERENS
O. BLANKE
Y. COPPENS


PIVS · IIII · PONTIFEX · OPTIMVS · MAXIMVS

Contents

Prologue
Antonio M. Battro, Stanislas Dehaene, Marcelo Sánchez Sorondo and Wolf J. Singer 12

Programme 14

List of Participants 17

Scientific Papers

Origins of Mind

Hominid Evolution and the Emergence of the Genus Homo
Yves Coppens 21

Human Origins from a Genomic Perspective
Svante Pääbo 36

Mind and Soul? Two Notions in the Light of Contemporary Philosophy
Enrico Berti 41

The Dynamic Brain and Consciousness

The Neuronal Correlate of Consciousness: Unity in Time Rather than Space?
Wolf Singer 51

Brain Rhythms for Cognition and Consciousness
Earl K. Miller and Timothy J. Buschman 68

The Brain Mechanisms of Conscious Access and Introspection
Stanislas Dehaene 79

Consciousness and Self-Consciousness in Favour of a Pragmatic Dualism in the Philosophy of Mind
Jürgen Mittelstraß 97

Neuroscience of Self-Consciousness and Subjectivity
Olaf Blanke 106

Towards a Neuroscientific Understanding of Free Will

Neural Mechanisms Underlying Human Choice in the Frontal Cortex
Tim Behrens 121

False Perceptions & False Beliefs: Understanding Schizophrenia
Chris D. Frith & Karl J. Friston 134

Addiction: A Disease of Self-Control
Nora D. Volkow and Ruben Baler 149

Understanding Layers: From Neuroscience to Human Responsibility
Michael S. Gazzaniga 156

Self-Knowledge and the Adaptive Unconscious
Timothy D. Wilson 170

Seven Ways Neuroscience Aids Law
Owen D. Jones 181

Interaction Between Two Readings: The Naturalistic and the Socratic "Know Thyself"
Marcelo Sánchez Sorondo 195

Sources of Human Comprehension and Incomprehension

Are There Innate Mechanisms That Make Us Social Beings?
Uta Frith 215

Natural Cooperation
Martin A. Nowak 237

Spiritualité de l'âme
Georges Card. Cottier, O.P. 241

Developmental Sources of Social Divisions
Elizabeth S. Spelke 250

Can Neuroscience Improve the Brain and Mind?

How Genes and Experience Shape the Human Will
Michael I. Posner, Mary K. Rothbart, Pascale Voelker & Yi-Yuan Tang 273

New Interfaces for the Brain
John Donoghue 287

***Homo Docens* and the Teaching Brain**
Antonio M. Battro 296

Final Statement
Neurosciences and the Human Person: New Perspectives on Human Activities
Georges M.M. Card. Cottier, Silvia Arber, Antonio M. Battro, Timothy Behrens, Enrico Berti, Olaf Blanke, Thierry Boon Falleur, Yves Coppens, Stanislas Dehaene, Christopher D. Frith, Uta Frith, Earl K. Miller, Jürgen Mittelstraß, Martin Nowak, Svante Pääbo, H.E. Msgr. Marcelo Sánchez Sorondo, Wolf J. Singer, Nora D. Volkow 305

Tables 315

Prologue

Working Group 8-10 November 2012 – Each generation of neuroscientists, philosophers and theologians has the task of analyzing and assessing new advances in the understanding of the human person, dignity and value within nature.

The Pontifical Academy of Sciences acknowledged this responsibility through two important workshops which should be seen as milestones.

The first was held in 1964 and was on 'The Brain and Conscious Experience' (Study Week, 28 Sept.-3 Oct. 1964, SV 30. Chairman: John C. Eccles; edited by Pietro Salviucci).

The second was held in 1988 and was on 'Brain Research and the Mind-Body Problem. Epistemological and Metaphysical Issues' (5 Oct. 1988, SV 79. Round Table Discussion. Chairman: Carlos Chagas; edited by Giuseppe del Re).

In 1964 the President of the Academy was George Lemaître and the Chancellor was Pietro Salviucci. There were 24 participants, including the Academicians John Eccles, Corneille Heymans, Giuseppe Moruzzi and Roger Sperry. Among the invited scientists there were Edgar Lord Adrian, Creutzfeldt, Ragnar Granit, Hebert Jasper, Benjamin Libet, Vernon Mountcastle and Hans-Lukas Teuber.

In 1988 the President was Carlos Chagas and the Directors were Enrico di Rovasenda and Renato Dardozzi. There were 22 participants. The theologians were Enrico di Rovasenda, Jan Schotte, and Giovanni Marchesi and the philosophers were Giuseppe del Re, Vittorio Mathieu, Peter Henrici, Paolo A. Rossi and Francesco Calvo. Among the scientists were the Academicians John C. Eccles, Jérôme Lejeune and János Szentagothai. This Round Table was preceded by a Study Week Session on the 'Principles of Design and Operation of the Brain' which had a long list of distinguished participants.

Continuing this important lineage of research and reflection, it is thus appropriate to gather for another meeting to continue the dialogue of the current generation between neuroscientists, philosophers and theologians in this scientific age. Significantly, this workshop on 'Neurosciences and the Human Person: New Perspectives on Human Activities' comes after another 24 years, the lifetime of a generation and the same time span that separated the first two meetings of the Academy.

Neuroscientists have made fundamental improvements since the last meeting in 1988 with the introduction of advanced neurobiological and

genetic technologies – and a corresponding new language – which deserve analysis in order to have a better understanding of the status of the human being that is in line with these new scientific discoveries. Philosophers and theologians, in their turn, are increasingly aware of the particular discoveries, epistemologies and languages that science has developed and try to interpret this new significant data in the light of the Socratic principle 'know yourself'. It follows that man's knowledge is not derived from a single perspective – that of external observation, explanation, and experimentation: this knowledge develops in the interface between the observation of nature and reflective understanding. The human being is an observable entity, like all organisms but at the same time it reflects on itself, it is a 'self-interpreting being'. Thus, understanding the human condition requires analysis of the various levels of knowledge and descriptions involving reconciliation between insights derived from the first, second and third person perspective in this age of rapid scientific progress. We hope that this dialogue between the different approaches and languages, which we propose for these three days, will enrich the contemporary understanding of the human person.

We may imagine another meeting of our Academy on this same issue in 2036 but we certainly cannot predict the topics and the technologies that will be discussed then. Our fields are expanding rapidly and the scientific, philosophical and theological challenges will increase accordingly.

We sincerely thank all the participants for their valuable contribution.

Antonio M. Battro, Stanislas Dehaene, Marcelo Sánchez Sorondo and Wolf J. Singer

Programme

Thursday, 8 November 2012

9:00	*Welcome* (Marcelo Sánchez Sorondo) *Introduction* (Antonio M. Battro)
	1. ORIGINS OF MIND Moderator: Uta Frith • Rapporteur: Wolf J. Singer
9:10	*Hominid Evolution and the Emergence of the Genus Homo!* **Yves Coppens**
9:30	Discussion
9:50	*Human Origins From a Genomic Perspective* **Svante Pääbo**
10:10	Discussion
10:20	*Mind or Soul? Two Notions in the Light of Contemporary Philosophy* **Enrico Berti**
10:40	Discussion
11:00	**Papal Audience**
12:30	Lunch at the Casina Pio IV
14:00	**General discussion session 1** (guided by Wolf J. Singer)
	2. THE DYNAMIC BRAIN AND CONSCIOUSNESS Moderator: Michael S. Gazzaniga • Rapporteur: Yves Coppens
15:00	*The Unity of Experience: Temporal Coherence Rather than Spatial Convergence?* **Wolf J. Singer**
15:20	Discussion
15:40	*Brain Rhythms, Cognition, and Consciousness* **Earl K. Miller**
16:00	Discussion
16:20	*Searching for Brain Mechanism of Conscious Access and Introspection* **Stanislas Dehaene**
16:40	Discussion
17:00	Coffee break
17:30	*Consciousness and Self-Consciousness. In Favour of a Pragmatic Dualism in the Philosophy of Mind* **Jürgen Mittelstraß**
17:50	Discussion
18:10	*Neuroscience of Self-Consciousness and Subjectivity* **Olaf Blanke**
18:30	Discussion
18:50	**General discussion session 2** (guided by Yves Coppens)
20:00	Dinner at the Casina Pio IV

Friday, 9 November 2012

3. TOWARDS A NEUROSCIENTIFIC UNDERSTANDING OF FREE WILL
Moderator: Michael I. Posner • Rapporteur: Jürgen Mittelstraß

9:00	*The Neuroscience of Human Choice* **Timothy E. Behrens**
9:20	Discussion
9:40	*False Perceptions and False Beliefs: Understanding Schizophrenia* **Christopher D. Frith**
10:00	Discussion
10:30	Coffee break
11:00	*Addiction: A Brain Disease of Free Will* **Nora D. Volkow**
11:20	Discussion
11:40	*Understanding Layers: From Neuroscience to Human Responsibility* **Michael S. Gazzaniga**
12:00	Discussion
13:00	Lunch at the Casina Pio IV
15:00	*Self-Knowledge and the Adaptive Unconscious* **Timothy D. Wilson**
15:20	Discussion
15:40	*Law, Neurosciences and Behaviour* **Owen D. Jones**
16:00	Discussion
16:20	*Interaction Between Two Readings: The Naturalistic and the Socratic "Know Thyself"* **Marcelo Sánchez Sorondo**
16:40	Discussion
17:00	Coffee break
17:30	**General discussion session 3** (guided by Jürgen Mittelstraß)
20:00	Dinner at the Casina Pio IV

Saturday, 10 November 2012

	4. SOURCES OF HUMAN COMPREHENSION AND INCOMPREHENSION Moderator: Nora D. Volkow • Rapporteur: Antonio M. Battro
9:00	*Are There Innate Mechanisms That Make Us Social Beings?* **Uta Frith**
9:20	Discussion
9:40	*The Evolution of Cooperation* **Martin A. Nowak**
10:10	Discussion
10:30	Coffee break
11:00	*The Christian View of the Human Person and the Soul* **Georges M.M. Cardinal Cottier**
11:20	Discussion
11:40	*Developmental Sources of Prejudice* **Elizabeth S. Spelke**
12:00	Discussion
12:20	**General discussion session 4** (guided by Antonio M. Battro)
13:15	Lunch at the Casina Pio IV
	5. CAN NEUROSCIENCE IMPROVE THE BRAIN AND MIND? Moderator: Elizabeth S. Spelke • Rapporteur: Stanislas Dehaene
15:00	*The Birth of the Mind* **Jacques Mehler**
15:20	Discussion
15:40	*How Genes and Experience Shape Will* **Michael I. Posner**
16:00	Discussion
16:20	*New Interfaces for the Brain* **John P. Donoghue**
16:40	Discussion
17:00	Coffee break
17:30	*Circuits for Action Diversification* **Silvia Arber**
17:50	Discussion
18:10	*Homo Docens and the Teaching Brain* **Antonio M. Battro**
18:30	Discussion
18:50	**General discussion session 5** and **Final Statement** (guided by Stanislas Dehaene)
20:00	Dinner at the Casina Pio IV

List of Participants

Prof. Werner Arber, President
Biozentrum, Department of Microbiology,
University of Basel
Basel (Switzerland)

H.E. Msgr. Marcelo Sánchez Sorondo, Chancellor
The Pontifical Academy of Sciences
(Vatican City)

Prof. Dr. Silvia Arber
Biozentrum, University of Basel
Department of Cell Biology
and Friedrich Miescher Institute,
Basel (Switzerland)

Prof. Antonio M. Battro
Academia Nacional de Educación
Buenos Aires (Argentina)

Prof. Timothy E. Behrens
Oxford Centre for Functional MRI of the Brain, FMRIB Centre,
John Radcliffe Hospital,
Oxford (UK)

Prof. Enrico Berti
Università degli Studi di Padova
Dipartimento di Filosofia
Padova (Italy)

Prof. Olaf Blanke
Bertarelli Foundation Chair in Cognitive Neuroprosthetics, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne;
1015 Lausanne (Switzerland)

Prof. Yves Coppens
Collège de France
Paris (France)

H.E. Card. Georges M.M. Cottier
Palazzo Apostolico
(Vatican City)

Prof. Stanislas Dehaene
Inserm-CEA, Cognitive Neuroimaging Unit,
CEA/SAC/DSV/DRM/NeuroSpin
Gif-sur-Yvette (France)

Prof. John P. Donoghue
Brown Institute for Brain Science
Brown University
Providence, RI (USA)

Prof. Christopher D. Frith
Leopold Müller Functional Imaging Laboratory, Wellcome Trust Centre for NeuroImaging, University College London (UK)

Prof. Uta Frith
Emeritus Professor of Cognitive Development, UCL Institute of Cognitive Neuroscience
London (UK)

Prof. Michael S. Gazzaniga
Sage Center for the Study of Mind
University of California, Santa Barbara
Santa Barbara, CA (USA)

Prof. Owen D. Jones
Macarthur Foundation Research Network on Law and Neuroscience
Vanderbilt University
Nashville, TN (USA)

Prof. Jacques Mehler
Scuola Internazionale Superiore
di Studi Avanzati (SISSA)
Trieste (Italy)

Prof. Earl K. Miller
The Picower Institute for Learning and
Memory, Department of Brain and
Cognitive Sciences, Massachussets
Institute of Technology
Cambridge, MA (USA)

Prof. Jürgen Mittelstraß
University of Constance
Center for Philosophy of Science
Constance (Germany)

Prof. Martin A. Nowak
Program for Evolutionary Dynamics,
Harvard University
Cambridge, MA 02138 (USA)

Prof. Svante Pääbo
MPI-EVA – Max Planck Institute
for Evolutionary Anthropology
D-04103 Leipzig (Germany)

Prof. Michael I. Posner
Department of Psychology
Institute of Cognitive and Decision
Sciences, University of Oregon
Eugene, OR (USA)

Prof. Wolf J. Singer
Max-Planck-Institute for Brain Research
Frankfurt am Main (Germany)

Prof. Elizabeth S. Spelke
Laboratory for Developmental Studies
Department of Psychology, Harvard
University
Cambridge, MA (USA)

Prof. Nora D. Volkow
National Institute on Drug Abuse, NIH
Bethesda, MD (USA)

Prof. Timothy D. Wilson
Department of Psychology
University of Virginia P.O. Box 400400
Charlottesville,VA 22904-4400 (USA)

Scientific Papers

Origins of Mind

Hominid evolution and the emergence of the genus *Homo*

Yves Coppens

I am very happy and much honoured to have been invited, for the third time, to this famous Academy, for a new working group.

If I understood well, my duty, here, is to give you the state of art of palaeoanthropology, the current way of understanding, with bones and teeth, the history of Man, of his close relatives and of his closest ancestors. I will try to do that.

⋆⋆⋆

Let us remember, for the pleasure, that Man is a living being, an eucaryot, a metazoaires, a chordate, a vertebrate, a gnathostom, a sarcopterygian, a tetrapod, an amniot, a synapsid, a mammal, a primate, an Haplorhinian, a Simiiform, a Catarrhinian, an Hominoidea, an Hominidae, an Homininae and that life, on earth, is around 4 billion years old, metazoaires, 2 billion years old, vertebrates, 535 million years old, gnathostoms, 420 million years old, mammals, 230 million years old, Primates, 70 million years old, Hominoidea, 50 million years old, Hominidae, 10 million years old.

And let us remember also that Primates, adapted to arboricolism and frugivory, developed three flourishing branches worldwide all over the tropics: the Plesiadapiforms, the Strepsirhinians composed of Adapiforms and Lemuriforms, and the Haplorhinians, composed of Tarsiiforms and Simiiforms, and that the Hominoidea are a superfamily of the Simiiforms, born in Eastern Asia, fifty million years ago, as I mentioned above. The Hominidae are a family of the Hominoidea, born in tropical Africa ten million years ago, and they include the last common ancestors of two subfamilies: the Homininae, us, and the Paninae, the Chimpanzees.

⋆⋆⋆

Ten million years for the Homininae-Paninae divergence may not be the right figure but it is an easy one, probably not too far from the real one. The debate about this date has always existed, especially between palaeontologists and geneticists. The palaeontologists used to prefer long chronologies (the Early Divergence Hypothesis), the geneticists, short ones (the Late Divergence Hypothesis). I remember that it had already been the reason

for the organization in May 1982 of the Pontifical Academy of Sciences' working group called *Recent Advances in the Evolution of Primates* (Chagas, 1983), reason that I suggested to the President Carlos Chagas.

Everybody agrees on the common ancestry of *Homo* and *Pan*; with the discovery of *Sahelanthropus*, 7 million years old, in Chad, palaeontologists are currently thinking of a divergence not too long before this fossil, considered a Homininae. Geneticists were suggesting less first (Scally *et al.*, 2012), and then more, up to 12 million years ago (Langergraber *et al.*, 2012). Svante Pääbo will tell us more about that.

So here let us accept 10 million years as a sort of average.

⋆⋆⋆

The progeny of these common ancestors, as we said, split into two main branches, Homininae and Paninae, probably because of the emergence of two environments, different enough, one more covered with more trees, one less, with less trees. The analysis of the diet (teeth) of several mammals of these upper Miocene levels shows, as a matter of fact, an increase in C4 plants, meaning a development of grasses (Cerling *et al.*, 2010; Uno *et al.*, 2011).

The subpopulation which happened to be in the more covered environment became the Paninae, the prechimpanzees and the chimpanzees – knuckle-walking locomotion and frugivory; the subpopulation which happened to be in the less covered environment became the Homininae, the prehumans and the humans – erect posture, double locomotion, bipedality and arboreality, and diet from trees (fruits) and from the ground (roots).

⋆⋆⋆

I. Early Prehumans

Let us take the "Homininae road". Between 10 and 4 million years ago, 3 genera and 4 species document the first step of this Prehumanity: *Sahelanthropus tchadensis*, 7 million years old, found in Chad, signed by Michel Brunet and 37 authors in 2002; *Orrorin tugenensis*, 6 million years old, found in Kenya, signed by Brigitte Senut and 5 authors in 2001; *Ardipithecus kadabba*, possibly 5.8 to 5.6 million years old, named by Yohannnes Haile Selassie in 2001; and *Ardipithecus ramidus*, 4.4 million years old, named by Tim White and 2 authors in 1994, both found in Ethiopia. Let us call them the early Prehumans.

They all share an erect posture; two of them, *Orrorin tugenensis* and *Ardipithecus ramidus* elegantly demonstrate, through their anatomy, the double locomotion, bipedality and arboreality, we previously mentioned.

The femora of *Orrorin* (3 have been found) show, for instance, apomorphic features with humans, elongated antero-posteriorly compressed

femoral neck with asymmetric distribution of cortex, spherical, anteriorly twisted head, shallow superior notch and developed gluteal tuberosity, precursor of the *linea aspera*. And in contrast, the humeral shaft (one has been found) shows an insertion of the *brachioradialis* muscle as a strong vertical crest characteristic of arboreal habits.

Demonstrating the same double locomotion, the pelvis of *Ardipithecus ramidus* is modified in its upper part to walk and run in odd contrast with its lower part still adapted to climb – powerful hip and thigh musculature; his foot still possesses an "os peroneum", known in monkeys but lost in Paninae, amazingly used here to help it walk in increasing its rigidity – a typical exaptation – in contrast with an abducted (grasping) big toe; its hand still looks very strong, able to support the weight of the body – palmigrady – but shows at the same time a good grip.

"*Ardipithecus* was at home both moving along trees on its palms and walking upright on the ground" wrote Brooks Hanson in the 2009 *Science* issue dedicated to the description of *Ardipithecus ramidus* (White *et al.*, 2009). It is obvious that this new environment and the adaptative answers in diet, posture and locomotion that the Homininae found to survive, had consequences in the organization of their brains and of course in their behaviour: regular food carrying, pair-bonding and reproductive crypsis (females did not advertise ovulation), suggests Owen Lovejoy in the same *Science* issue.

According to the very low degree of sexual dimorphism, readable in *Ardipithecus ramidus*, it is clear that these early Prehumans were still living in quite a covered environment, synonymous of protection.

⋆⋆⋆

II. Classic Prehumans

Around 4 million years ago, the climate seems to change again, in the same dryer direction. It looks more like an increase in the same change, having happened around 10 million years ago. The landscape is opening: its covered part, which still existed, is diminishing and its opened part is obviously increasing.

I suggested the existence of that change many years ago because, having studied the Proboscidians, I was surprised by the fact that in Kanapoi, in Kenya, 4 million years ago, *Elephas ekorensis* and *Mammuthus subplanifrons*, the Elephantidae eating grasses, were appearing whereas *Anancus kenyensis, Stegotetralodon orbus* and *Primelephas gomphotheroides* eating leaves (and existing in Lukeino, 6 million years ago for instance) had disappeared. And there are now some beautiful confirmations of these observations with the demonstration of the strong development of C4 plants at these geological

times (Lee-Thorp *et al.*, 2012) (Lukeino is one of the sites of *Orrorin tugenensis*, Kanapoi is one of the sites of *Australopithecus anamensis*).

So this new period, between 4 and 3 million years ago, is the time of the Australopithecines sensu lato, as well as the Kenyanthropines, who were more aggressive, walked better, and had started eating meat.

We know 2 genera and five or six species of these Prehumans that we could call classic Prehumans to differentiate them from the early Prehumans. They come from Chad, Ethiopia, Kenya, Tanzania and South Africa, a sort of concentric tropical circle around the equatorial forest.

In chronological order, they are:

- *Australopithecus anamensis* from Kenya and Ethiopia, starting around 4 million years ago, signed by Meave Leakey and three coauthors in 1995;
- *Australopithecus bahrelghazi*, from Chad, 3.5 million years old, signed by Michel Brunet and five other authors in 1996;
- *Australopithecus afarensis*, starting around 3.9 million years ago from Tanzania and Ethiopia and signed by Donald Johanson and two authors in 1978 (which can now be reduced to its Tanzanian part, because the type chosen was a mandible from Laetoli in Tanzania, the Ethiopian part having recently become *Australopithecus chamensis* (not yet completely accepted);
- *Australopithecus prometheus*, from South Africa (previously known by the nickname of Little Foot) around 3 million years old; this old name, given by Raymond Dart (Dart, 1948) to some South African specimens, has been recently proposed by Ronald Clarke, to name it Little Foot (Clarke, 1995, 2012);
- *Kenyanthropus playtops*, 3.5 million years old from Kenya, described by Meave Leakey and six coauthors in 2001, which could be synonymous of *Australopithecus bahrelghazi* and which is modern looking thanks to its orthognathic flat midface.

Australopithecus afarensis and *Australopithecus chamensis* from East Africa as well as *Australopithecus prometheus* from South Africa show the double locomotion that we described in *Orrorin tugenensis* and *Ardipithecus ramidus*, but a double locomotion not completely similar to the one of the early Prehumans. The orientation of the lesser trochanter of the femora is more posterior in *Australopithecus*, for instance, than in *Orrorin*, where it is medially projected; the head of the femora has the same diameter as its neck in *Australopithecus*, but is much larger in *Orrorin*, the head is twisted *posteriorly* in *Australopithecus*, but anteriorly in *Orrorin* and so on. The interpretation of

their differences is that Australopithecines walk and run more efficiently.

And for the first time, one species, *Australopithecus anamensis,* showed, through a particularly stable hind limb (knee joint) and a quite instable forelimb (elbow joint), exclusive bipedality. I wonder whether this specificity so important in anatomy and its consequent behaviour would not necessitate a different generic name.

In comparison with the early Prehumans, the size of these classic Prehumans is about the same or slightly increasing, but their sexual dimorphism is completely different; it is very important indeed, reflecting a much more open environment, consistent with what we said about the fauna, the flora, and the anatomy of these Homininae.

As far as the teeth are concerned, these classic Prehumans seem to have chosen two diets, two adaptations, both then possible, in showing an increasing or decreasing size of the post canine teeth. *Australopithecus afarensis* and *Australopithecus chamensis*, for instance, have chosen to increase the size of these teeth; *Kenyanthropus playtops*, to decrease it.

Furthermore, after the description of cut marks on a few bones, in a site, Dikika, 3.4 million years old, where only one species of Homininae has been discovered so far, *Australopithecus afarensis* (*chamensis*?), the idea that some of these classic Prehumans were already partly carnivores, which means more omnivorous, has been claimed and more or less accepted (McPherron *et al.*, 2010).

⋆ ⋆ ⋆

In summary, between 10 and 3 million years ago (chapter I and chapter II) a subfamily, the Homininae, was born in central and east Africa, and evolved in central, east and south Africa, because of climate changes. It is currently documented by 5 genera, *Sahelanthropus, Orrorin, Ardipithecus, Australopithecus, Kenyanthropus* and 10 species. All these Homininae were, as we said, tropical and African without any exception, permanently upright, walking and climbing first before becoming exclusively biped, with a slowly increasing endocranial capacity, 300 to 350cc in *Ardipithecus ramidus*, 400cc in *Australopithecus chamensis*, and complexity (more convolutions and better irrigation), a slow reduction of their prognathism at different speeds and with a trend to reduce or to increase the size of their cheek teeth. The diversity of these Prehumans, as far as locomotion, dentition, consequent behaviour and diets are concerned, is important and fascinating; it is obvious that we will find more fossils and greater diversity.

⋆ ⋆ ⋆

III. Late Prehumans and Early Humans

Around 3 million years ago, probably a bit less, climate changed again in the dryer direction, having started 10 million years ago and having increased 4 million years ago; but this time global cooling and tropical drought were severe: less and less trees, more and more grasses and the need for everyone to find solutions to adapt to these new conditions to survive.

Global cooling appeared in the study of oxygen isotope ratios O^{18}/O^{16}, in the tests of microorganisms collected in deep-sea cores in the Atlantic and in the Indian oceans. Tropical drought, now very well-known and studied, appeared first in the sediments of the lower Omo river basin in Ethiopia (Coppens, 1975, 1978 a and b, 1983a, b and c, 1985; Boisserie *et al.*, 2008), because these sediments are the only ones in tropical Africa to offer a clear continuous, very fossiliferous and thick enough (more than one kilometre) deposit of these geological times, between a little more than 3 million years at the bottom to a little less than 1 million years at the top. Among many examples of the Omo sequence documenting this climate change, let us take only two of them, one from the fauna, and one from the flora.

As far as the faunal example is concerned, I have chosen to give you the quantification of two tribes of Antelopes, the Tragelaphini, living in open forests, covered areas and more or less thick bush, and the Alcelaphini, adapted to run in open countries, without too much water. In the lower levels, Tragelaphini represented 33% of the Antelopes, Alcelaphini 9%; in the upper levels, Tragelaphini are 3% of the Antelopes, Alcelaphini, 29%.

Let us now mention the figures obtained by palynology; an index of a number of pollens of trees on a number of pollens of grasses has been done in lower and upper levels; for the earliest levels, this index reached the figure of 0.4; for the upper levels, the same index got the figure of 0.01.

I hope that this demonstration of climate change, through these two examples, has been convincing.

The Homininae have been giving three brilliant answers to this crisis. Let us call these answers a robust one and two gracile ones.

The robust one can be schematically called a physical answer: bigger body size, more massively built, impressive new masticatory equipment for chewing vegetarian fibrous diet, but only small, allometric, development of the brain; we know 2 or 3 genera and 4 species to document this answer.

The gracile ones can be very schematically called an intellectual answer in East Africa: much bigger brain and omnivorous diet but small body, and a more mobile answer in South Africa, better pelvis for better bipedality but small brain; we know 2 genera and 4 species to document these second and third answers.

This pack of 4 or 5 genera and 8 species can be called late Prehumans and early (or first) Humans, the Person.

The robust answer looks particularly interesting at the same time because of its homogeneity, its diversity and the limits of this diversity. By homogeneity, I mean that the three answers we will describe have found the same strategy: robust body, robust cheek teeth and small brain. By geographical and ecological diversity, I mean that in three biogeographical and ecological niches, the answers are, as a matter of fact, comparable but not similar.

In the Afar area (east Africa) the robust form is *Australopithecus garhi* (2.5 million years old), long hind limbs but long forelimbs as well, very big teeth (canine but anterior also), prognathic face and small brain (450cc), discovered and published by Berhane Asfaw and five other authors in 1999.

In eastern Africa sensu lato, south of Ethiopia, Kenya, Tanzania, Malawi, we are dealing with a robust lineage, *Paranthropus* or *Zinjanthropus aethiopicus* – that Arambourg and myself discovered and described in 1967 – and *Paranthropus* or *Zinjanthropus boisei* – that Louis Leakey described in 1959 – the first 2.7 to 2.3, the second 2.3 to 1.2 million years old. Both are very robust forms, with a prognathic dish midface in *Paranthropus aethiopicus*, much less in *Paranthropus boisei*, a shallow palate in *P. aethiopicus*, much deeper in *P. boisei*, a small cranial capacity (400 cc) in *P. aethiopicus*, larger (530cc) in *P. boisei*, and a very specialized dentition, very small anterior cutting teeth in a straight line and very large post canine grinding cheek in two almost straight rows, with very thick enamel.

In South Africa, a similar form of specialized Prehuman, *Paranthropus robustus* was described by Robert Broom as soon as 1938; its characteristics are about the same as for the east African robust parade: strongly built body, skull with robust superstructures, like a sagittal crest, wide dish face, deep postorbital constriction, small anterior, strong posterior teeth, deep palate, small brain (around 500cc).

This diversity is a beautiful example of adaptation but also a beautiful example of parallel adaptation as well; it seems that these solutions were found independently by very close but different lineages. I would not be surprised if the origin of *Paranthropus* (or *Zinjanthropus) aethiopicus-boisei* were *Australopithecus afarensis* (*chamensis*?), and the origin of *Paranthropus robustus* were *Australopithecus prometheus* (Little Foot).

The gracile solutions, contrasting with the robust one, are heterogeneous according to the ecogeographical niches where they had to express themselves.

One of these solutions, "found" by South African Prehumans, was a strategy of more efficient mobility; it has been documented by a lineage of *Australopithecus africanus*, *Australopithecus sediba*. *Australopithecus africanus* is famous

because it was the very first species of Prehuman ever recognized, described and named by Raymond Dart in 1925. *Australopithecus africanus*, which could well be around 2.4-2.5 million years old (at Sterkfontein), seemed to have fore limbs still adapted to climb but hind limb already fully adapted to walk; it has a globular skull with a moderate to marked alveolar prognathism, small endocranial capacity (440cc) and a dentition with relatively small incisors and canines and relatively large premolars and molars.

Australopithecus sediba recently found and published by Lee Burger and six authors in 2010 is 1.9 to 2 million years old – it still possessed long and powerful forelimbs to climb but a derived hand with a long thumb to grip, a primitive foot but a derived wider pelvis, a human-like sacrum and strong femora, synonymous of a good bipedality associated with a more evolved face but a still very small cranial capacity (420cc).

And the second solution, "found" by the east African Prehumans, was a strategy to survive in an environment probably dryer than the South African one, an obvious bigger reorganized brain and a new dentition for a clear omnivorous diet, where meat had become a part of new feeding habits. This solution is documented by a new genus, *Homo*, and two possible species, *Homo habilis* and *Homo rudolfensis*, (Leakey *et al.*, 1964, Groves *et al.*, 1975).

The consequences of this natural event and natural adaptation to this event by natural selection, have been fantastic; more meat, more animal protein, means better brain; more brain means a new level of thought, curiosity, new approaches of life, cognitive, intellectual, spiritual, ethic, aesthetic, new behaviours.

I am conscious that the words that I am employing are philosophically and scientifically too provocative, too strong and at last inappropriate, even wrong. But for a palaeontologist, a field palaeontologist, after years of surveys and excavations, discovering in the middle of an obvious dramatic climate crisis the very first stone-made tools and their makers, is just fantastic. Suddenly you are in front of the first Human, the true Human being, in front of a Person, capable of anticipating enough to create a shape for his own future use or pleasure.

It is to recall the pioneer role of the lower Omo river sequence in Ethiopia in the discovery of the correlation between the 2.7 climate change and the emergence of the genus *Homo* that many years ago I proposed the name of (H)Omo event, with an H in brackets to link *Homo* and Omo (in a very bad pun).

And scientifically speaking 2.6 or 2.7 is the date of the discovery by one, or by several Homininae. It is, as a matter of fact, not currently possible to claim for sure who is, or who are, the makers of second-degree stone tools.

⋆ ⋆ ⋆

In summary, between 3 and 2 million years ago, because of a change in climate, classic Prehumans invented three brilliant solutions to survive: the robust one (late Prehumans) which would last almost 2 million years and the gracile ones, one of them being more mobile and the other being Man, still alive almost 3 million years later (first Humans). With the first Humans (at least) emerges a new level of consciousness, probably never reached before, giving rise to the very first manmade artefacts planned according to their projected function.

It is the emergence of the Person. As soon as the genus *Homo*, the human genus, was born, Man was there, complete, even if he has been evolving during the 2 and a half million years after his birth and even if he is still expected to evolve in the future.

The answers to the questions of where, when, how and why did a Prehuman become a Man, could be:

In tropical Africa (maybe only East Africa);

Between 3 and 2 million years ago, around 2.7;

In developing his brain and changing his diet and his dentition and, of course, his behaviour to try to adapt to the dramatic climate change he had to cope with.

⋆⋆⋆

IV. Classic and late Humans

The genus *Homo*, being omnivorous, which means carnivorous pro parte, enlarged his territory; a carnivore always has a larger ecological niche than any herbivore.

But being carnivorous, the genus *Homo* had to hunt and consequently became more mobile.

Having a bigger brain, more plicated, with a better irrigation, the genus *Homo* became consciously organized to explore more territories for hunting and gathering and maybe also for curiosity.

As we said before, the genus *Homo* made tools, invented shapes for functions or not, and as soon as he had done that, kept doing it forever. Man and tools became a couple, no tools without Man, no Man without tools. And since making a tool, as soon as the first one, is a symbolic gesture, I would say "no symbol without Man and no Man without symbol".

As *Homo*'s adaptation to climate change was a success, his population probably increased demographically, very slowly, of course, but at a speed fast enough to be obliged to move, to extend his territory.

In summary, being more mobile because of his new diet, more curious because of his new brain, better equipped because of the tools he made,

more numerous because of his adaptative success, Man, the first Man, the first species of the genus *Homo*, moved.

And some environmental reasons could probably be added to these previously mentioned ones – a natural extension of his ecological niche – to support the idea of a very early movement of the genus *Homo*.

For environmental, biological and cultural reasons it was the very first species of the genus *Homo* who moved, extending his territory, as soon as 2 to 2.5 million years ago, almost anywhere and everywhere (with a latitudinal climatic limit) in the Old World, Africa, Europe and Asia.

And since, with *Homo*, there are stone tools, it becomes easier to trace his movements.

I would briefly like to list some data to support the idea of a very early in and out of Africa n°1 movement of *Homo* as soon as *Homo habilis*.

Africa:	more than 2 million years, in Algeria, Aïn Boucherit (tools);
Middle East:	more than 2 million years in Israel, Yiron (tools); 1.8 million years in Georgia, Dmanissi (bones and tools);
Europe:	1.6 in Italy, Pirro Nord (tools); 1.6 in France, Lezignan (tools); 1.2 in Spain, Sima del Elefante (bones and tools), Barranco León, Fuente Nueva 3 (tools).
Asia:	1.9 in Pakistan, Riwat (tools); more than 2 million years in India, Masol (tools); 1.8 in Malaysia, Lunggong (tools); 1.6 in Indonesia, Sangiran, Modjokerto (bones and tools); 1.7 to 2 in China, Majuangou, Yuanmou, Longuppo, Renzidong, Longuddong (tools).

Then, it seems that there is:

1) A *Homo habilis, Homo ergaster, Homo erectus* lineage, all over this huge area. But as *Homo* was not demographically numerous enough to exchange genes everywhere all the time, the very first species of the genus *Homo* could have become the second and the third, but not the fourth because new subspecies or species of *Homo* emerged by isolation (by sea or by ice);

2) A probable *Homo antecessor, Homo heidelbergensis, Homo neandertalensis* lineage in Europe and then, later, in the Middle East and Central Asia;

3) A population of Siberia, the Denisovans, derived from the *Homo neandertalensis* lineage, discovered by geneticists (Sante Pääbo and his staff), remaining a Siberian spot from a much larger territory (Asia);

4) A probable endemic lineage of *Homo erectus* in the Indonesian islands, *Homo erectus* and *Homo soloensis* in Java*, Homo floresiensis* in Flores;

5) A probable *Homo erectus*, *Homo rhodesiensis, Homo sapiens* lineage in Africa and the Middle East (Morocco, Israel);

6) Another probable *Homo erectus* lineage "evolved *Homo erectus*" in the Far East (*Homo sapiens* like) (China);

7) And then a possible emergence of Homo *sapiens sapiens* in Africa around 200,000 years ago and a possible movement, for environmental reasons, of this subspecies out of Africa again (n°2), through the Bab el Mandeb and the Sinai around 100,000 years ago (Eriksson *et al.*, 2012), interbreeding with the populations previously established in Asia.

Homo sapiens sapiens has been found in:

Middle East:	100,000 years in Israel, El Zuttiyeh (bones);
Asia:	100,000 years in China, Zhirendong (bones) (Liu *et al.*, 2010);
	75,000 years old in India, Narmada valley (bone) (Sankhyan *et al.*, 2012);
	74,000 years old in Malaysia, Lenggong (tools) (Zuraina Majid, personal communication, 2012);
	60,000 years old in Laos, Tam Pa Ling (bones) (Demeter *et al.*, 2012).

And it was apparently this population of *Homo sapiens sapiens* who moved again to Europe 50,000 years ago, to Siberia 30 to 40,000 years ago, to Java 50 to 60,000 years ago and then to Flores 10,000 years ago.

And *Homo neandertalensis*, Java Man, Denisovan and Flores Man became extinct, from 30,000 years ago to 10,000 years ago, but not without leaving some "souvenirs" that the geneticists are trying, more and more successfully, to recognize and identify.

⋆ ⋆ ⋆

I must say that I am feeling less comfortable in the systematic world of the genus *Homo*. I am not sure that all these species – *Homo ergaster, Homo antecessor, Homo heidelbergensis, Homo rhodesiensis* etc. – that we mentioned and *Homo cepranensis, Homo georgicus* etc. that we did not mention, really exist, or at least exist in the same way as *Sahelanthropus tchadensis, Australopithecus anamensis* or *Paranthropus boisei*. It is not impossible that culture has had a retroaction on biology and that the numerous human species of the same genus *Homo* are only grades with permanent interbreeding potentialities except, maybe, for isolated and specialized forms like *Homo neandertalensis* (the later one) or the tiny *Homo floresiensis*.

I must say as well that, if I am absolutely convinced by the first "out of Africa" n°1 and by its antiquity – 2.5 to 2 million years ago – I am not completely convinced by the second, 100,000 years ago. I think that, if the second does exist, it would not be the second but maybe the tenth or the hundredth; as soon as the movements of people from north-eastern Africa to the Middle East became climatically and environmentally possible, I can-

not understand why the human population would have stopped passing, in both directions, after moving once.

The closing of Europe because of glaciation is understandable; the existence of hand-axes in the tool kit, 1.7 million years ago in Africa, and only 700,000 years ago in Europe, for instance, is important data for sure. The existence of an upper Pleistocene climate change pushing *Homo sapiens sapiens* out of Africa 100,000 years ago is also a good datum. But between "out of Africa" n°1 and "out of Africa" n°2, I guess there were several "out of Africas" as well as "out of Asias".

⋆⋆⋆

Homo sapiens sapiens got to Australia by boat around 40,000 years ago, to America through the almost empty Behring straight, by foot or by boat, around 30,000 years ago (at least) and to Greenland, by foot, 5,000 years ago, and then, obviously by boat to Melanesia, Micronesia, Polynesia, some thousands of years ago.

So since 10,000 years ago there is only one species and one subspecies of Homininae on the Earth, *Homo sapiens sapiens*.

We will have to wait for the peopling of other planets to be able to get, by long enough isolation and genetic drift, new human subspecies or species and new bunches of Homininae…

⋆⋆⋆

In summary, Humans extended their territory very early beyond Africa through the Sinai and the Bab el Mandeb roads to the whole of Eurasia; but the peopling was too small for too large an area to stay genetically stable and started to create a generous specific diversity. But a new subspecies, probably born in Africa, extended its territory through the same roads to Eurasia and then to the whole world; this subspecies, *Homo sapiens sapiens*, being obviously dominant everywhere, all the previous human species became extinct.

⋆⋆⋆

This conclusion deals with the history in time and space of the subfamily Homininae, the subfamily, zoologically speaking, we belong to. It is a long history ten million years old, starting in tropical Africa by an odd adaptation to a new behaviour because of climate change, upright posture, and continuing, still in tropical Africa, by another adaptation to another behaviour because of another climate change: a "better" brain.

This succession of natural events and of adaptations has been the natural reason for the emergence of a new being, the human genus, developing in

natural environments a new environment, the cultural one, and bringing with it a new consideration of the individual, the Person.

I would like to conclude with a reaction by one of my own grandmothers, who told me, without any possible discussion, "If you, my grandson, descend from the Apes, I, your grandmother, do not".

She was wrong, as far as the natural history of man was concerned but she was right in defending the dignity of the Person.

Bibliography

Arambourg, C., Coppens, T. 1967, Sur la découverte dans le Pléistocène inférieur de la vallée de l'Omo (Ethiopie) d'une mandibule d'Australopithécien, *C. R. Acad. Sc.*, 265, 589-590.

Asfaw, B., White, T., Lovejoy, O., Latimer, B., Simpson, S. & Suwa, G. 1999, *Australopithecus garhi*: A New Species of Early Hominid from Ethiopia, *Science* 284, 629-635

Berger L.R., Ruiter D.J. de, Churchill S.E., Schmid P., Carlson K.J., Dirks P.H.G.M., Kibii J.M., 2010, *Australopithecus sediba*: a new species of *Homo*-Like Australopith from South Africa, *Science*, 326, 195-204.

Boisserie J.-R,, Guy F., Delagnes A., Hlukso L.J., Bibi F., Yonas Beyene Y., Guillemot C., 2008, New palaeoanthropological research in the Plio-Pleistocene Omo Group, Lower Omo Valley, SNNPR (Southern Nations, Nationalities and People Regions), Ethiopia, *C R Palevol*, 7, 429-439.

Broom, R. 1938, The Pleistocene Anthropoid apes of South Africa, *Nature*, 142: 377-379.

Broom, R., Robinson R.T., 1952, Swartkrans ape-man, *Paranthropus robustus*, *Transvaal Museum Mem.*, 6:1-123.

Brunet, M., Beauvilain, A., Coppens, Y., Heintz, E., Montaye, A.H.E. & Pilbeam, D. 1996, *Australopithecus bahrelghazali*, a new species of early hominid from Koro Toro region (Chad). *C. R. Acad. Sc.*, 322, 907-913.

Brunet, M., Guy, F., Pilbeam, D, Mackaye, H.T., Likius, A., Ahounta, D., Beauvilain, A., Blondel, C., Bocherens, H., Boisserie, J.-R., Bonis, L. De, Coppens, Y., Dejax, J., Denys, C., Duringer, P., Eisenmann, V., Fanone, G., Fronty, P., Geraads, D., Lehmann, T., Lihoreau, F., Louchart, A., Mahamat, A., Merceron, G., Mouchelin, G., Otero, O., Pelaez-Campomanes, P., Ponce De León, M., Rage, J.-C., Sapanet, M., Schuster, M., Sudre, J., Tassy, P., Valentin, X., Vignaud, P., Viriot, L., Zazzo, A. & Zollikofer, C. 2002, A new Hominid from the Upper Miocene of Chad, Central Africa, *Nature*, 418, 145-151.

Cerling T.E., Harris J.M., Leakey M.G., Passey B.H., Levin N.E., 2010, Stable carbon and oxygen isotopes in East African mammals: modern and fossil, *Cenozoic Mammals of Africa*, Werdelin L., Sanders W.J. eds., University of California Press, Berkeley, 941-952.

Chagas C. ed. 1983, Working group on *Recent advances in the Evolution of Primates*, 24-27 mai 1982, Pontificiae Academiae Scientiarum, *Scripta Varia* 50, 204 pages. Online at www.pas.va/content/accademia/en/publications/scriptavaria/evolutionofprimates.html

Clarke R.J., Tobias P.V. 1995, Sterkfontein Member 2 foot bones of the oldest South African hominid, *Science*, 269: 521-524.

Clarke, R.J., 2012, The history of research in human evolution in Africa and what lessons have been learned, *World Heritage papers*, 33, Heach 2, Unesco, 44-68.

Coppens Y., 1975, Evolution des Hominidés et de leur environnement au cours du

Plio-Pléistocène dans la basse vallée de l'Omo en Ethiopie. *C. R. Acad. Sc.*, 281, 1693-1696.

Coppens Y., 1978 a, Evolution of the Hominids and of their environment during the Plio-Pleistocene in the lower Omo Valley, Ethiopia, Bishop W.W. ed., *Geological Background to fossil Man,* Londres, 499-506.

Coppens Y., 1978 b, Les Hominidés du Pliocène et du Pléistocène d'Ethiopie, chronologie, systématique, environnement, Piveteau J. ed., *Les origines humaines et les époques de l'intelligence*, Masson ed., Paris, 79-106.

Coppens Y., 1983 a, Les plus anciens fossiles d'Hominidae, Chagas C. ed., *Recent Advances in the Evolution of Primates*, Pontificiae Academiae Scientiarum *Scripta Varia*, 50, 1-9.

Coppens Y., 1983 b, Les Hominidés du Pliocène et du Pléistocène d'Afrique orientale et leur environnement. Table ronde "Morphologie évolutive, morphogénèse du crâne et anthropogénèse", VIIIe Congrès de la Société Primatologique Internationale, Paris, 155-168.

Coppens Y., 1983 c, Systématique, phylogénie, environnement et culture des Australopithèques ; hypothèses et synthèse, *Bull et Mém. Soc. Anthrop. Paris*, XIII, 10 (3), 273-284.

Coppens Y., 1983 d, *Le singe, l'Afrique et l'Homme*, Fayard, 152 pages.

Coppens Y. ed., 1985, *L'Environnement des Hominidés au Plio-Pléistocène*, Masson 468 pages.

Dart, R., 1925, *Australopithecus africanus*, the man-ape of South Africa, *Nature*, 115, 195-199.

Dart, R. 1948, The Makapansgat Proto-Human *Australopithecus Prometheus*, *Am. J. Phys. Anthrop.*, 6,259-284.

Demeter F., Shackelford L.L., Bacon A.-M., Duringer P., Westaway K., Sayavongkhamdy T., Braga J., Sichanthongtip P., Khamdalavong P., Ponche J.-L., Wang H., Lundstrom C., Patole-Edoumba E., Karpoff A.-M. 2012, Anatomically modern human in Southeast Asia (Laos) by 46 ka., *PNAS*, 109, 14375-14380.

Eriksson A., Betti L., Andrew D. Friend A.D., Lycett S.J., Singarayer J.S., von Cramon-Taubadel N., Valdes P.J., Balloux F., Manic A., 2012, Late Pleistocene climate change and the global expansion of anatomically modern humans, *PNAS*.

Groves C.P. and Mazak V., 1975, An approach to the taxonomy of the *Hominidae*; gracile Villafranchian Hominids of Africa, *Casopis pro Mineralogi Geologi*, 20, 225-246.

Haile-Selassie, Y. 2001, Late Miocene hominids from the Middle Awash, Ethiopia, *Nature*, 412: 178-181.

Johanson, D.C., White, T.D., Coppens, Y. 1978, A new species of the genus *Australopithecus* (Primates: Hominidae) from the Pliocene of eastern Africa, *Kirtlandia*, Cleveland, 28, 1-14.

Langergraber K.E., Prüfer K., Rowney C., Boesch C., Crockford C., Fawcett K., Inoue E., Inoue-Muruyama M., Mitani J.C., Muller M.N., Robbins M.M., Schubert G., Stoinski T.S., Viola B., Watts D., Wittig R.M., Wrangham R.W., Zuberbühler, Pääbo S., Vigilant L., 2012, Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution, *PNAS,* 109 (39) 15716-15721.

Leakey, L.S.B., 1959, A new fossil skull from Olduvai, *Nature*, 184, 491-493.

Leakey L.S.B., Tobias P.V., Napier J.R. 1964, A new species of the genus *Homo* from Olduvai Gorge, *Nature*, 202, 7-9.

Leakey M.G., Feibel C.S., Mcdougall I., Walker A. 1995, New four million-years-old hominid species from Kanapoi and Allia Bay, Kenya, *Nature* 375: 565-571.

Leakey M.G., Spoor F., Brown, F.H., Gothogo P.N., Kiairie Ch., Leakey L.N., Mcdougall I., 2001, New hominin genus from eastern Africa shows diverse middle Pliocene lineages, *Nature* 410, 433-439.

Lee-Thorp J., Likius A., Mackaye H.T., Vignaud P., Sponheimer M., Brunet M., 2012, Isotopic evidence for an early shift to C_4 resources by Pliocene hominins in Chad, *PNAS*.

Liu W., Jin C.Z., Zhang Y.Q., Cai Y.J., Xing S., Wu X.J., Cheng H., Edwards R.L., Pan W.S., Qin D.G., An Z.S., Trinkaus E., Wu X.Z., 2010, Human remains from Zhirendong, South China, and modern human emergence in East Asia, *PNAS*, 107, 45, 19201-19206.

McPherron S.P., Alemseged Z., Marean C.W., Wynn J.G., Reed D., Geraads D., Bobe R., Bearat H.A. 2010. Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia, *Nature*, 466:857-860.

Sankhyan A.R., Badam G.L., Dewangan L.N., Chakraborty S., Prabha S., Kundu S., Chakravarty R., 2012, New Postcranial Hominin Fossils from the Central Narmada Valley, India, *Advances in Anthropology*, 2, 125-131

Scally A., Dutheil J.Y., Hillier L.W., Jordan G.E., Goodhead I., Herrero J., Hobolth A., Lappalainen T., Mailund T., Marques-Bonet T., McCarthy S., Montgomery S.H., Schwalie P.C., Tang Y.A., Ward M.C., Xue Y., Yngvadottir B., Alkan C., Andersen L.N., Ayub Q., Ball E.V., Beal K., Bradley B.J., Chen Y., Clee C.M., Fitzgerald S., Graves T.A., Gu Y., Heath P., Heger A., Karakoc E., Kolb-Kokocinski A., Laird G.K., Lunter G., Meader S., Mort M., Mullikin J.C., Munch K., O'Connor T.D., Phillips A.D., Prado-Martinez J., Rogers A.S., Sajjadian S., Schmidt D., Shaw K., Simpson J.T., Peter D. Stenson, Turner D.J., Vigilant L., Vilella A.J., Whitener W., Zhu B., Cooper D.N., Jong P.de, Dermitzakis E.T., Eichler E.E., Flicek P., Goldman N., Mundy N.I., Ning Z., Odom D.T., Ponting C.P., Quail M.A., Ryder O.A., Searle S.M., Warren W.C., Wilson R.K., Schierup M.H., Rogers J., Tyler-Smith C., Durbin R., 2012, Insights into hominid evolution from the gorilla genome sequence, *Nature*, 483, 169-175.

Senut B., Pickford M., Gommery D., Mein P., Cheboi K., Coppens Y. 2001, First Hominid from the Miocene (Lukeino formation, Kenya), *C. R. Acad Sc.*, 332, 137-144.

Uno K.T.; Cerling T.E., Harris J.M., Kunimatsu Y., Leakey M.G., Nakatsukasa M., Nakaya H., 2011, Late Miocene to Pliocene carbon isotope record of differential diet change among East African herbivores, *PNAS*, 108, 6509-6514.

White T.D. Asfaw B., Beyene Y., Haile Selassie Y, Lovejoy C.O., Suwa G., Wolde-Gabirel G., 2006, *Ardipithecus ramidus* and the Paleobiology of Early Hominids, *Science*, 326, 75-86.

White T.D., WoldeGabriel G., Asfaw B, Ambrose S.H., Yonas Beyene Y., Raymond L Bernor R.L., Boisserie J.R., Currie B., Gilbert H., Haile Selassie Y, Hart W.K., Hlusko L.J., Clark Howell F., Kono, R.K. Lehmann T., Louchart A., Lovejoy, C.O., Renne P.R., Saegusa H., Vrba E.S., Wesselman H., Suwa G., 2006, Asa Issie, Aramis and the origin of *Australopithecus*, *Nature*, 440, 883-889.

White T.D. Asfaw B., Beyene Y., Haile Selassie Y, Lovejoy C.O., Suwa G., Wolde-Gabirel G., 2009, *Science*, 326, special section, 11 articles, 60-107.

White, T.D., Suwa, G., Asfaw, B. 1994, *Australopithecus ramidus*, a new species of early hominid from Aramis, Ethiopia, *Nature*, 371, 306-312.

Human Origins from a Genomic Perspective

Svante Pääbo[1]

Humans with skeletons indistinguishable or almost indistinguishable from those of present-day humans appear for the first time in the fossil record of Africa between 100,000 and 200,000 years ago. These "anatomically modern humans" then appear outside Africa shortly before 100,000 years ago, and shortly before 50,000 years ago start to spread across Eurasia and the rest of the world. By that point, their behavior is in several respects radically different from that of earlier forms of humans that had existed in Africa for several million years and in Eurasia for about two million years. For example, while earlier forms of humans had made much the same sorts of stone tools for hundred of thousands of years, modern human technology changed rapidly, such that stone tools become different in different geographical regions. Art in a form that present-day humans intuitively recognize as art appears only after modern humans had appeared. And modern humans start spreading across the entire world by crossing even bodies of water where land is not visible on the other side. With one possible exception, this had never been done before.

These new human behaviors that appear with modern humans reflect cultural developments that set present-day humans apart from all other primates and have allowed them to become extremely numerous, to populate areas of the world where they could not survive without technology, and eventually to dominate parts of the biosphere. A fundamental question in modern biology is to understand the genetic underpinnings of these changes.

In order to begin to do this it is necessary to compare the genomes of present-day humans to those of our closest relatives, so-called "archaic humans", who are not "modern humans" in the sense that they did not share these behaviors. The closest and best-known relatives of modern humans are the Neandertals, who appear in the fossil record of Europe about 300,000 or 400,000 years ago and live in western Eurasia until they become extinct about 30,000 years ago. Over the past thirty years my laboratory

[1] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany.

has developed techniques that have led to the recent determination of DNA sequences from the entire genome of Neandertals (Green *et al.*, 2012) as well as from another closely related group of extinct humans, the Denisovans, in southern Siberia (Reich *et al.*, 2010; Meyer *et al.*, 2012). This allows us to begin to address two fundamental questions with regard to the origins of modern humans, which I briefly outline below.

Genetic traces of archaic humans in people today

A question debated among paleontologists for decades is whether modern humans mixed genetically with archaic humans when they spread across Africa and Eurasia or whether modern humans replaced archaic humans without any mixture.

Since genetic variation in Africa is greater than in the entire rest of the world and most genetic variants that exist outside Africa are very similar to variants found inside Africa, genetic anthropologists had generally inferred that a total replacement of all archaic humans by modern humans had occurred, even if some were of a different opinion (e.g. Wall *et al.*, 2009). However, when the Neandertal genome was sequenced (Green *et al.*, 2012), it was found that Neandertals shared slightly more genetic variants with present-day people outside Africa than with people inside Africa. This is best explained by a scenario in which modern humans, when they emerged out of Africa, mixed with Neandertals, perhaps in the Middle East, and then carried a genetic contribution from Neandertals along with them when they spread across the world. They then passed this contribution on to their children such that on the order of 1-4 percent of the genomes of people outside Africa today stem from Neandertals. Recently, the size distribution of the segments of Neandertal DNA in present-day people has been used to date the mixing of modern humans and Neandertals to sometime between 40,000 and 90,000 years ago (Sankararaman *et al.*, 2012).

Using similar techniques of DNA retrieval from ancient bones, the genome of a Denisovan, a relative of Neandertals, was determined from the Altai Mountains in southern Siberia (Reich *et al.*, 2010; Meyer *et al.*, 2012). When this genome was compared to those of present-day humans it was found that they have contributed on the order of 5 percent of the genomes of people that now live in Papua New Guinea and other parts of Melanesia. In addition, it has been shown that people in eastern Asia carry slightly more Neandertal DNA sequences than people in Europe (Meyer *et al.*, 2012). It is therefore possible that mixing of modern humans and Neandertals occurred not only in the Middle East but also later as modern humans spread across Eurasia. In addition, patterns of variation in Africa have

been interpreted to mean that also in Africa, other groups of archaic humans mixed genetically with modern humans (Hammer *et al.*, 2011).

It is therefore clear that present-day humans carry a direct genetic contribution from earlier extinct forms of humans. This contribution is found scattered in pieces across the genome of any one individual. Each such piece exists in a few percent of people today, and in a single individual they add up to a few percent of the genome. The rest of the genome of any single non-African individuals, well over 90 percent, originates within the past 200,000 years in Africa where the transformation to modern human behavior and anatomy occurred. Some of the pieces of the genome that come from Neandertals and other archaic humans may contribute to physiological differences among people today, for example in how the immune system functions (Abi-Rached *et al.*, 2011), but most of these variants are likely to have no functional consequences whatsoever. Nevertheless, it is of interest that present-day humans carry genetic contributions from earlier forms of humans whom they encountered as they spread across the globe. When we study the genetic origins of modern humans it is therefore appropriate to use the plural form, *origins*, as different parts of our genome have different origins.

The genetic basis of the modern human condition

Of fundamental importance for understanding, from a biological perspective, what sets modern humans apart from earlier forms of humans is to identify all genetic changes that are shared in identical form among all or almost all humans today but where our closest evolutionary relatives, such as the Neandertals and the Denisovans, shared other variants with the apes and other primates. These are genetic changes which together define modern humans as a group as distinct from our closest extinct relatives as well as all other primates. They are, in a sense, a "genetic recipe" for being a modern human.

The recent determination of a complete Denisovan genome (Meyer *et al.*, 2012) has allowed the compilation of a list of almost all such changes. It contains all positions in the genome where all or almost all humans today, no matter where on the planet we live, are identical but where the Denisovan is identical to the apes. Interestingly, this list is not extremely long. It contains 111,812 single nucleotide changes among the approximately 3 billion nucleotides that make up the entire human genome. It also contains 9,499 insertions and deletions of a number of adjacent nucleotides. These changes go back to mutations that occurred in modern human ancestors after their separation from the ancestors of Neandertals and Denisovans some 400,000 to 600,000 years ago and before perhaps 50,000 years ago,

after which the dispersal of humans across the planet made it impossible for any genetic change to spread to all humans.

A major challenge for human biology in the next decade is to investigate which of these changes have functional consequences and to understand how they have contributed to the unique cultural developments that have characterized the last 50,000 years of human history. How could this be achieved?

Investigations of genetic features unique to modern humans

The investigation of biological features unique to humans is not a trivial task since the very fact that these changes are unique to humans means that no obvious animal models are available. In spite of this, I believe that three approaches are possible.

First, the human genome is small enough and the number of new mutations that occur in each newborn baby large enough (~50-100) that all mutations compatible with human life exist in the current world population. However, most of them are very rare. In the future, when new technologies will make it possible to sequence the genomes of millions of people it will become possible to identify even rare variants that represent mutations back to an earlier, ancestral state. The physiological consequences of these mutations can then be studied.

Second, induced pluripotent stem cells are cells derived from adult tissues that can be induced to become, for example, nerve cells or liver cells in the laboratory. In such cells, DNA sequences can be "back-mutated" to the ancestral state and their effects on cellular functions can be studied in different cell types under different conditions. This will also make it possible to combine several, and eventually many, such changes affecting a certain organ system or biochemical pathway.

Third, to study the effects of mutations in a living organism, human mutations can be introduced into a model organism such as the mouse and aspects of human-specific traits studied.

Each of these approaches has obvious limitations. For example, they are limited to the identification or introduction of one or a few mutations whereas many mutations may exhibit effects only in concert with other mutations. Similarly, many mutations may not function in the same way in the laboratory mouse as in human beings due to differences in the genetic make-up between mice and humans. Nevertheless, the study of this list of human-specific genetic changes is an undertaking worth pursuing because an understanding of what constitutes the biological basis for modern humans, especially in the realm of cognition, is of fundamental importance

for understanding how modern humans came to explode in population size and dominate the biosphere. It may also provide an additional inroad into an understanding of diseases that affect cognitive traits that are unique to modern humans, such as language delay, speech disorders, or autism.

References

Abi-Rached L, *et al.* (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* 334, 89-94.

Green RE, *et al.* (2010) A draft sequence of the Neandertal genome. *Science* 328:710-722.

Meyer M, *et al.* (2012) A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science.* doi: 10.1126/science.1224344.

Reich D, *et al.* (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468(7327):1053-1060.

Sankararaman, S. *et al.* (2012) The date of interbreeding between Neandertals and modern humans. *PLoS Genetics* [Epub ahead of print].

Hammer MF *et al.* (2011) Genetic evidence for archaic admixture in Africa. *Proc. Natl Acad Sci U S A.* 108, 15123-8.

Wall JD, *et al.* (2009) Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.* 26, 1823-7.

Mind and Soul? Two Notions in the Light of Contemporary Philosophy

Enrico Berti

Are mind and soul two separate entities? The question is legitimate because of the meaning that is commonly given to terms "mind" and "soul". The term "mind" is mostly used in modern scientific and philosophical language to indicate an entity, which some say is not really distinct from the body (monists), but is reduced to the body (materialists) or otherwise resolves all bodies in itself (idealists), while according to others it is really distinct from the body, forming a second world distinct from that of body (dualistic) or even causes a third world, distinct from both the body and the mind (Popper). The term "soul" is used in traditional philosophical language, and especially in religious language, to indicate a separate entity from the body, which in some cases even pre-exists the body, becoming incarnate from time to time in different bodies, and in any case survives the body, going toward prizes or punishments after the death of the body. Both of these meanings, in my opinion, are deviant, precisely because they lead to believe that mind and soul are entities, namely entities existing in themselves, in the same way that bodies are. The real problem therefore is not whether mind and soul are two distinct entities, but first of all whether they truly are existing entities in the same manner as the bodies.

The term "mind" and the term "soul" both derive from the ancient Greek word *psychê*, which is at the basis of a whole family of other terms commonly used in science, philosophy and everyday language, such as "psychic", "psychology", "psychiatry", which have the advantage of not alluding to existing entities in the same manner as bodies, but rather indicate properties, or dispositions, or behaviours, or processes, in short, phenomena without an existence of their own but belonging, so to speak, to subjects that are generally human beings or even animals, and therefore bodies.

However, the term *psychê* in ancient Greek was used with very different meanings, basically two, the first of which is at the root of both the common concept of "mind" and the common concept of "soul", while the second is at the root of the family of words that start with "psycho". The first meaning is the one that has had more luck in the history of Western culture. It dates back to the Orphic religion, whereby *psychê* is a "demon", a sort of intermediate being between man and god, which pre-exists the body and

with the birth of man (or animal) is incarnated, namely enters the body, lives with it in a more or less conflictual relationship throughout the life of the body and then leaves the body upon the latter's death, to be incarnated in a different body of man or animal better or worse depending on its merits or its faults, or continues to live independently from any body in a world better or worse than this, again depending on its merits or its faults.

This concept of *psychê* was taken up in antiquity by philosophers such as Pythagoras and in particular Plato, who determined its immense fortune. From Plato, in fact, it passed on to Christianity, where it gave rise to the dualism between "spirit" and "flesh" and the dogma of the immortality of the soul, decreed by the Church in the Renaissance (sixteenth century), which has remained until now in the Catechism of the Catholic Church. However, it found an extremely effective philosophical formulation in the early modern period with Descartes, one of the fathers of modern philosophy and science, who, to save the dogma of the immortality of the soul, conceived it as a substance, whose essence consists of thought, entirely independent of the body and capable of acting, that is living, independently of it and surviving it: the *res cogitans*. *Psychê* thus took on, already in the antiquity, the Latin name of *anima,* which means "soul", and, in the case of man, where the soul proves capable of performing functions exceeding those of the other animals, that of *mens*, meaning "mind".

The Cartesian dualism of body and soul, or mind and body, however, was immediately exposed to numerous and extremely serious difficulties, such as justifying the causal actions performed by the body in relation to the soul (feelings, emotions, desires) and those carried out by the soul with regard to the body (movements, alterations), so the Cartesian demonstration of the substantiality of the soul was judged (by Kant) a "paralogism", namely an erroneous reasoning, and the dualism of body and soul was solved either in a materialistic monism (Lamettrie, Holbach, and others) or in an idealistic monism (Berkeley, Fichte, Bradley, Gentile and others). In the philosophy of the twentieth century it gave rise to the so-called *Mind-Body Problem*, on which there is endless literature both from the philosophical side, namely by the so-called "philosophy of the mind", and by the scientific part by the so-called Artificial Intelligence, "cognitive sciences", or the neurosciences.

The most vigorous denunciation of Cartesian dualism was made in the twentieth century by Gilbert Ryle, editor the oldest philosophical review in the English language, not surprisingly called *Mind*, and author of an important book entitled *The Concept of Mind* (1949). In its critical part the book destroys what the author calls the Cartesian "dogma of the ghost in the machine", showing that it arises from a "category error" of common

language, namely that of placing the "mind" in the same category of entities to which they belong the bodies. In its constructive part, however, it conceives the mind as a set of "dispositions" or "skills", which give rise to a series of objectively controllable behaviours. For this reason Ryle's concept was mistaken for a form of behaviourism, and his book – as Daniel Dennett, one of the scholars of the current philosophy of the mind, acknowledged in the preface to the new edition – had long been underestimated, while it is still fully relevant to this day.

What is not yet recognized is the fact that Ryle put forward again without saying so (no philosopher is cited in his book), the other major concept of *psychê* developed in antiquity, the one contained in Aristotle's *De anima* (in Greek *peri psychês*), according to which *psychê* is not a substance distinct from the body, as the Orphists, the Pythagoreans and Plato would have liked, but is the "form" or "'first act" of an organic body, that is of a body that is formed by organs, equipped with life in potency, that is, capable of living (*De anima* II 1). Therefore – says Aristotle – it makes no sense to say that the soul feels pain and joy, courage and fear, gets angry, feels and thinks. This in fact would be like saying that the soul can weave or build a house. We must say instead that man feels compassion, learns or thinks through the soul (ibid., I 4). That Ryle was Aristotelian, as was John L. Austin, another great representative of the Oxford School, teacher of the current philosopher of mind John Searle, is now admitted by all, because Oxonian philosophy as an analysis of the common language did nothing but take the logic and ontology of Aristotle (in this case his doctrine of categories). But the importance of the Aristotelian conception of *psychê* has been recognized by those "philosophers of mind" that explicitly referred to it, interpreting it first in a physicalistic key, namely as a reduction of psychic phenomena to physical phenomena (Feigl, Slakey, Matson), then in a functionalistic key, that is, by assimilating the *psychê* to a computer program (Fodor, Dennet and the first Putnam), and finally in a correctly hylomorphistic key, that is, as the form of a living body (Martha Nussbaum and the last Putnam, in *Word and Life*, 1995). Finally it was embraced by the *Catechism of the Catholic Church* (Vatican City 1993), which refers explicitly to the Aristotelian definition of the soul as *forma corporis* (§ 365), accepted by the Council of Vienne (1312).

To understand how this could happen it is necessary to understand exactly the concepts of "form" and "first act". Form is not a metaphysical entity, as those who give the term "metaphysics" the meaning of "mysterious", "hidden" and "misleading" believe, but it is the way a certain matter is organized and functions. For example, a sensible form visible to the naked eye is the wheel, which allows a certain matter (stone, wood, metal, rubber)

to roll and thus perform a series of functions that would not be possible without it (this is neurobiologist Roger Sperry's example). An intelligible form, one that is understandable by means of a concept, is the formula of water (*formula* in fact means "little form"), H_2O, which is not one of the components of the molecule of water (two hydrogen atoms and one oxygen atom), but without which there would be no water. Now, neither the form of the wheel nor the formula of water are matter and yet exist, denying, therefore, any form of monism, materialistic or idealistic, but they are not entities in themselves, and therefore deny any form of dualism.

Even clearer is the notion of "first act", which Aristotle explains with the example of possessing a science, say geometry, which is different from using it (second act), for example by demonstrating a theorem. As the first act of a body having life in potency, the soul is therefore the actual possession of the ability to live in a body that serves as a tool, where the subject of life is neither the soul nor the body, but the living being itself. Aristotle says that if an axe were a living being, its soul would be the ability to cleave, and if the eye were an animal, its soul would be sight. An eye without sight, like a painted eye or a stone eye, is not a true eye, but only an eye "by homonymy", that is in name only, like a dead man is a man in name only. The soul in short is simply what makes the difference between a living organism, be it a human being, an animal or a plant, and a dead body, that is a corpse. In order for this soul to be present it is not necessary for the living being to act out the functions which the soul makes him capable of, but it is sufficient that they are present as a capacity, or – in the language of Scholastics – as "active potency", corresponding to the Aristotelian "first act".

Life is a series of different functions, of which the basic ones are feeding and reproducing. These functions also belongs to plants, which therefore, according to Aristotle, have a soul, the vegetative soul. Perceiving and moving, functions typical of animals, are added to the above functions and belong to the sensitive soul. Finally, we have thinking and willing, functions proper of human beings, which are added to the above and belong to the intellectual soul, i.e. the human soul. What distinguishes these various kinds of soul? Modern genetics has determined that the differences between genera and living species, and even between single individuals of the same species, depend on the DNA sequence, the acid content in the cell nucleus. Therefore, a famous biologist, Max Delbrück, Nobel laureate for medicine, wrote that Aristotle, with his notion of soul as form, that is as a programme, pioneered the discovery of DNA (in *Of Microbes and Life*, ed. by J. Monod-E. Borek, New York 1971, pp. 50-55). Indeed DNA is matter (i.e., composed of molecules containing proteins and I do not know what else), but the "sequence"

of its components, which distinguishes a plant from an animal and from a human being, and even a human individual from another, is a "formula", that is a form. Accepting this view, Hilary Putnam wrote that the soul understood in the Aristotelian sense is not just a programme comparable to computer software, because the programme is compositionally plastic, that is, it can be performed by various kinds of matter, while the soul is computationally plastic, that is, it has the ability to carry out different programmes.

Aristotle believed that the main organ in an animal, the one managing all of the soul's functions, was the heart, because he had seen that it was the first organ that was formed in the chick embryo (thus discovering what Harvey then called "epigenesis"), while modern science has shown that it is the brain and has succeeded in identifying some of the processes that occur in it, thanks to the neurosciences. The soul, understood in the Aristotelian sense, is the ability of the brain, or of the entire organism through the brain, to carry out these processes, ranging from the most basic functions, called physical, to the higher and more complex ones, called psychic. Indeed, the brain can perform its functions only if it is alive, that is, if its encephalogram is not flat: the soul, understood in the Aristotelian sense, is what distinguishes a living brain with a non-flat encephalogram from a dead brain with a flat encephalogram. It is then up to the neurosciences, or to the philosophy of mind, to understand and explain how to carry out these functions. The difference between so-called physical, or biological functions, and so-called psychic or mental functions, consists in the fact that the former are carried out by means of organs, that is, by bodily tools, and in this sense are material, while the latter do not require other organs – in addition of course to the brain to which they belong – and therefore are inherently immaterial. This was already recognized by Aristotle, when he said that the intellect does not take on material qualities, such as heat and cold, and has no organ (*De anima* III 4, 429 at 24-27), and was confirmed by Searle (*Mind: A Brief Introduction*, Oxford, 2004), who showed that conscience is not ontologically reducible to the brain, even though it is causally attributable to brain processes.

The problem raised at this point by the supporters of the soul is if a soul understood in this way, i.e. in the Aristotelian sense, can survive the body. The answer seems likely to be negative: indeed, for which subject would such a soul have the ability to carry out certain functions? And what functions could it carry out without the body if, according to Aristotle, even thought requires, as a starting point, the images provided by the senses, that is, by the body? Thomas Aquinas, who was Aristotelian, was very much aware of this when he said that the soul, after the death of the body, is no

longer a person (*S. Th.* 1, 29, 1 ad 5m; *Pot.* 9, 2, ad 14m, *C.G.* 4, 79), and when he stated, against the Averroists, that the subject of thought is the "individual man" (*hic homo*). The Christian faith, in its original formulation, suggests a belief in the resurrection of the whole person, body and soul, even before its belief in immortality understood as the survival of the soul separate from the body. Indeed, the Apostles' Creed recites, "I believe in the resurrection of the flesh" and the Nicene-Constantinopolitan Creed states, "I believe in the resurrection of the dead". The Church herself, in the prayer for the dead, says, "Eternal rest grant unto them O Lord", thus comparing the condition after death to sleep. The very canon of the Mass exhorts us to pray "for those who have fallen asleep in the hope of resurrection". But this is obviously a matter of faith, not philosophy or science.

Another problem arising from the results of the neurosciences is the existence of freedom, or free will, i.e. the question of whether our actions, or rather our decisions, are just the result of neurological processes that take place in our brain or if it is our will that causes them to happen. The consequences of the denial of freedom would in fact be disastrous for ethics, because they would destroy any moral responsibility, and there would be no difference between Socrates and Genghis Khan, or between Gandhi and Hitler. Benjamin Libet's experiments are often quoted in support of this denial, from which it appears that our awareness of a decision takes place with a delay of about 500 milliseconds compared to the neurological process that initiated it, which therefore appears to be unconscious, and therefore involuntary (*Mind Time: The Temporal Factor in Consciousness*, Cambridge, MA, 2004).

This too is an old problem that dates back to the interpretation of a passage of Aristotle (*Metaph.* VI 3), where the philosopher shows how certain events are the necessary result of a chain of actions and therefore are neither planned nor willed by the person who started the chain. The Stoics interpreted this phenomenon as "destiny" (*heimarmenê*), while Alexander of Aphrodisias in his *De fato* confuted them, showing that the passage of Aristotle acknowledges the concurrence, in determining the chain, of human initiative and a series of accidental causes that change its course. This is the ancient debate on "determinism", in which Cicero also participated in his *De fato* and which developed during the Renaissance with Pico della Mirandola (*Adversus astrologiam*) for example.

In fact Libet himself acknowledges to the conscious will the ability to permit or prohibit actions started at a subconscious level, that is, those that according to Aristotle are produced by desire (*orexis*) and according to Freud by *libido* (the inability to control desire was termed *akrasia* by Aristotle in book VII of the *Nicomachean Ethics*). In short, the subject of the action, the

one who decides and acts accordingly, is neither the subconscious nor conscience, neither desire nor will, but man. This was also repeatedly expressed by Thomas Aquinas who, following Aristotle, always indicated in the individual man (*hic homo*) he who not only thinks but also wills and loves (*De malo*, q. 6, art. un.).

The manuals of scholastic philosophy (see for example Sofia Vanni Rovighi, *Elementi di filosofia*, Milan, 1953, vol. IV) taught that freedom cannot be proved, because if it were provable, it would no longer be freedom, but necessity. Freedom is a matter of experience, such as the fact of experiencing pleasure or pain. We feel we are free, we have experienced the power to carry out or not carry out a certain action or to choose among different actions. A sign of this is what the English moralists (Shaftesbury, Hume) called "moral sense", such as the guilt and remorse that we feel when we think we have done a bad deed, and the satisfaction we feel in the opposite case. British psychologist Richard Holton (*Willing, Wanting, Waiting*, Oxford 2009) has argued recently that moral sense proves the existence of the brain's free will. For him the experience of free will is demonstrated by the experience of forming decisions and keeping resolutions, which require an effort on the part of the agent.

The fact that the moral sense is the product of evolution, especially the evolution of social life, as claimed by Patricia Smith Churchland (*Braintrust: What Neurosciences Tell Us About Morality*, Princeton, 2009), based on the study of what happens in the primate brain, does not mean that moral judgements have no value (for example, a judgement such as "torturing a child is a cruel action", according to Putnam, is a valid judgement, whatever its origin). Even Euclidean geometry, or Gödel's mathematics, are products of evolution, like all human activities, but that does not mean that the Pythagorean theorem or Gödel's incompleteness theorems are not valid. We must not confuse the origin of a proposition with its truth-value.

The Dynamic Brain and Consciousness

The Neuronal Correlate of Consciousness: Unity in Time Rather than Space?

Wolf Singer

The state of the art and explanatory gaps

For many decades the search for the neuronal underpinnings of cognitive and executive functions has been guided by the behaviorist view that the brain is essentially a highly complex and versatile stimulus response machine. Consequently, neuroscientists set out to study the responses of neurons to sensory stimulations across the various stages of the processing hierarchy, and analyzed activation patterns associated with motor output, hoping that these strategies would eventually lead to reductionist explanations of the neuronal mechanisms that support cognition, memory, decision making, planning and motor behavior. The strategy to follow the transformation of activity from the sensory surfaces over the numerous levels of hierarchically-organized processing structures to the respective effector organs proved to be extremely fruitful. Comparison of brains from different species provided compelling evidence that basic principles according to which neurons function and exchange signals have been preserved throughout evolution with only minor modifications. For the comparatively simple nervous systems of certain invertebrates this behaviorist approach allowed for near complete descriptions of the neuronal mechanisms underlying particular behavioral manifestations. This nurtured the expectation that pursuing this research strategy would sooner or later allow us to explain in the same way the more complex behavior of mammals and ultimately also the highly differentiated cognitive functions of primates and human subjects. However, over the last decades, the pursuit of the behaviorist approach accumulated evidence that requires a revision of the classical hypothesis that emphasizes serial feed-forward processing of sensory information within hierarchically organized architectures. Although the existence of processing streams devoted to the various sensory modalities is undisputed, comprehensive analysis of the brain's connectome revealed as the prevailing principles distributedness of functions and reciprocity of interactions. It has also become clear that the brain is by no means a stimulus driven system. Rather, it is self-active, permanently generating highly structured, high dimensional spatio-temporal activity patterns. These patterns are far from being random

but seem to reflect the specificities of the functional architecture that is determined by the genes, modified by experience throughout post-natal development and further shaped by learning. These self-generated activity patterns in turn seem to serve as priors with which incoming sensory signals are compared. Perception is now understood as an active, reconstructive process in which self-generated expectancies are compared with incoming sensory signals. The development of methods that allow simultaneous registration of the activity of large numbers of spatially distributed neurons revealed a mind-boggling complexity of interaction dynamics that eludes the capacity of conventional analytical tools and because of its non-linearity challenges hypotheses derived from intuition. These new and fascinating insights impose revision of concepts and unravel explanatory gaps that were not visible a few decades ago.

The rather detailed knowledge about the response properties of individual neurons in different brain structures is in harsh contrast to our ignorance of the complex and highly dynamic processes through which these neurons interact in order to produce specific behaviors. In-depth analysis of the brain's connectome revealed complex, nested small world architectures in which the principles of re-entry and distributedness prevail (Van den Heuvel and Sporns, 2011). Evidence from invasive and non-invasive multi-site recordings indicates that most higher brain functions result from the coordinated interaction of large numbers of neurons that become associated in a context and goal dependent way into ad-hoc formed functional networks that are dynamically configured on the backbone of the anatomical connections (for review see von der Malsburg *et al.*, 2010). Evidence also indicates that these interactions give rise to extremely complex spatio-temporal patterns that are characterized by oscillations in a large number of different frequency bands that can synchronize, exhibit phase shifts and even cross frequency coupling (Uhlhaas *et al.*, 2009). In the light of these novel data, the brain and in particular the neocortex appears as a self-active, self-organizing "complex system" that exhibits non-linear dynamics, is capable of utilizing multiple dimensions for coding (space, amplitude, phase), operates in a tightly controlled range of self-organized criticality (Shew *et al.*, 2009), (edge of chaos) and constantly generates highly structured, high dimensional activity patterns that are likely to represent stored information. However, how exactly information is encoded in the trajectories of these high dimensional and non-stationary time series is largely unknown and the subject of increasingly intense research. Moreover, with the exception of a few studies in which selective manipulation of the activity of defined neuron groups were shown to affect behavior in a particular way (Salzman

et al., 1992; Houweling and Brecht, 2008; Han *et al.*, 2011) most of the available evidence on the relations between neuronal responses and behavior is still correlative in nature. This makes it difficult to determine whether an observed variable is an epiphenomenon of a hidden underlying process or is causally involved in accomplishing a particular function. Thus, systems neuroscience now faces the tremendous challenge to analyze the principles of distributed dynamic coding and to obtain causal evidence for the functional role of specific activation patterns in order to distinguish between functionally relevant variables and epiphenomena. In conclusion, we have to abandon classical notions on the neuronal representation of perceptual objects. The consequence is that it became again unclear how the distributed processes that deal with the various properties of a perceptual object, its visual, haptic, acoustic, olfactory and gustatory features, are bound together in order to give rise to a coherent representation or percept. Given this, it may appear more than bold to attempt to identify the neuronal correlates of consciousness, the probably highest and most mysterious of our cognitive functions.

An attempt to define the explanandum

Most languages have coined a term for consciousness. Thus, it must be a robust phenomenon on which human beings can agree. However, while it is easy to use the term, it is virtually impossible to give a formal definition of what exactly it means. Nevertheless, the implicit understanding of what it is to be conscious seems to be sufficiently clear and widely accepted to justify search for its neuronal correlates and ultimately, to identify the neuronal mechanisms that enable a subject to be conscious of something. In their seminal paper, (Crick & Koch, 1990) proposed that consciousness is a specific cognitive function and as such must have neuronal correlates that can be analyzed with tools of the natural sciences. With the development of non-invasive imaging technologies, the tools became available to actually pursue this project and the search for the neuronal correlates of consciousness (NCC) has become a mainstream endeavor.

Before discussing some of the proposed theories for NCC, I shall attempt to give an operational definition of what I mean when referring to awareness and consciousness or in other terms what it means to be aware of something or to be conscious. Subjects will be considered as aware of something if they are able to report the presence or absence of the content of a cognitive process irrespective of whether this content is made available by recall from stored memories or drawn from actual sensory experience. Thus, the criterion for awareness is the reportability of the presence of a cognitive

content. These reports can in principle consist of any motor response but to be on the safe side, it is requested that the report be verbal. The reason is that behavioral responses can be obtained under forced choice conditions, that clearly indicate that the brain has processed and recognized the respective sensory material and produced a correct response even though the subject may not have been aware of having perceived the stimulus. There is thus an inherent ambiguity in non-verbal responses. They can but need not necessarily signal awareness which constrains research on the NCC in animal experiments. Since consciousness is so difficult to define an attempt will be made to avoid this term and rather use the adverb "consciously" and the adjective "conscious" in order to further specify particular brain states or aspects of a perceptual process. Also, no attempt will be made to address the hard problem of consciousness research, the problem to explain the phase transition from neuronal processes to the qualia of subjective experience (Chalmers, 2000).

The search for NCC needs to take into account a number of distinct properties that characterize the state of being aware of something. One important feature is unity or relatedness: Contents that one is aware of are experienced as simultaneously present and related with each other. Because of the distributed organization of brain processes, mechanisms supporting phenomenal awareness must be able to bind together computational results obtained in multiple specialized and widely distributed processing areas. Another feature of awareness is that the contents that one is aware of change continuously but are bound together in time, appearing as a seamless flow that is coherent in space and time. Finally, subjects are only aware of a small fraction of ongoing cognitive operations. Thus, there must be a mechanism that determines which signals subjects actually become aware of. As signals that subjects are not aware of are also readily processed and impact behavior (Dehaene *et al.*, 1998; Van Gaal *et al.*, 2008), these gating mechanisms must have a more subtle effect on neuronal interactions than simply blocking information processing. Therefore, the identification of the NCC requires analysis of the mechanisms that gate access to awareness.

Some competing hypotheses

One class of theories focuses on the philosophical implications of the hard problem without attempting to provide detailed descriptions of putative neuronal mechanisms (Searle, 1997; Metzinger, 2000; Dennett 1992; Chalmers, 2000). Solutions to the hard problem have also been sought for by transcending current concepts on neuronal processes and incorporating theories borrowed from other scientific disciplines. The most prominent of

these approaches assumes that phenomena unraveled by quantum physics also play a role in neuronal processes and might be able to account for the emergence of consciousness from the material interactions in the brain (Hameroff, 2006; Penrose, 1994). As none of the predictions of these theories are at present amenable to experimental verification, these will not be discussed further. Another class of theories pursues more modest goals and attempts to examine neuronal mechanisms potentially capable of supporting awareness of cognitive contents. Their aim is to define neuronal mechanisms supporting the unitary character of awareness, its coherence in space and time and the control of states distinguishing between conscious and unconscious processing.

The intuitively most plausible solution for the unity of awareness is convergence of the results obtained in distributed processing areas to a singular structure at the top of the processing hierarchy. Theories derived from this intuition predict the activation of specific cortical areas when subjects are aware of stimuli. Consequently, these regions should remain inactive during unconscious processing of the same material. Likewise, lesions of these putative areas should abolish the ability to become aware of perceptual objects. So far a region with such "observer functions" has not been identified and this option is considered theoretically implausible (Dennet, 1992). There is also little if any experimental evidence for such a scenario. Behavioral and brain imaging studies have shown that unconscious processing engages very much the same areas as conscious processing, including frontal and prefrontal cortex (Lau and Passingham, 2007; van Gaal *et al.*, 2008). Thus, there is no compelling evidence for specific areas supporting conscious processing. Lesions of the processing structures proper lead to a selective elimination from conscious perception of those aspects of the stimulus material that are processed in these regions but access to awareness is unimpaired for other contents. A prominent example for such conditions are agnosia and blindsight (Cowey & Stoerig, 1991). There are projection systems in the brain whose destruction abolishes all conscious experience but these cannot be considered as NCC. Rather, these systems adjust the activation level of the brain and are necessary for the maintenance of the state during which conscious processing of stimulus material is possible.

Another class of theories favors the notion that the mechanisms supporting awareness of stimulation material are distributed and do not require anatomical convergence. Baars (1997) and Dehaene *et al.* (2006) proposed that there is a work space of consciousness whose neuronal correlate is a widely distributed network of neurons located in the superficial layers of the cortical mantel. As mentioned above, these neurons are reciprocally

coupled through a dense network of cortico-cortical connections that have features of small-world networks. The proposal is that subjects become aware of signals if these are sufficiently salient to ignite coordinated activity within this workspace of consciousness. This is assumed to be the case for signals that either have high saliency because of high physical energy of the stimuli or because they are made salient due to attentional selection.

Yet another and related proposal is that subjects become aware of contents, irrespective of whether they are triggered by sensory events or recalled by imagery from stored memories, if the distributed neurons coding for these contents get organized into assemblies characterized by coherent, temporally structured activity patterns. In this case, the critical state variable distinguishing conscious from non-conscious processing would be the spatial extent and the precision of coherence of temporally structured neuronal responses (Rodriguez *et al.*, 1999; Metzinger, 2000; Varela *et al.*, 2001).

In the following evidence will be reviewed in support of the latter hypothesis. However, before discussing this evidence it is required to briefly recall the reasons why temporal coherence should matter in neuronal processing.

The formation of functional networks by temporal coordination

Because of the small world architecture of the cortical connectome any neuron can communicate with any other neuron either directly or via only a few interposed nodes. Thus, efficient and highly flexible mechanisms are required that permit selective routing of signals and assure that only those neurons effectively communicate with one another that need to interact in order to accomplish a particular task. Evidence from multisite invasive recordings and from non-invasive registration of global activity patterns with magneto-encephalography or functional magnetic resonance imaging indicates that indeed functional sub-networks are configured on the fly on the backbone of fixed anatomical connections in a task and goal dependent way. A mechanism that can accomplish such fast and selective association of neurons and gate neuronal interaction is the temporal coordination of oscillatory activity (Gray *et al.*, 1989; Fries, 2005). Since the discovery (Gray & Singer, 1989) that spatially distributed neurons in the primary visual cortex tend to engage in oscillatory responses in the beta and gamma frequency band when activated by appropriately configured contours and that these oscillatory responses can synchronize over large distances within and across cortical areas and even hemispheres, numerous studies have confirmed that oscillations and their synchronization in different frequency bands are an ubiquitous phenomenon in the mammalian brain. Pacemakers of these oscillations are reciprocal interactions in local networks of inhibitory and ex-

citatory neurons. The long distance synchronization of this oscillatory activity appears to be achieved by several mechanisms operating in parallel: Long range excitatory cortico-cortical connections, long range inhibitory projections and pathways ascending from nuclei in the thalamus and the basal forebrain (for review see Uhlhaas *et al.*, 2009). When neurons engage in oscillatory activity, they pass through alternating cycles of high and low excitability. At the peak of an oscillation cycle neurons are depolarized, highly susceptible to excitatory input and capable of emitting action potentials. In the subsequent trough of the cycle the membrane potential is hyperpolarized and membrane conductance is high because of strong GABAergic inhibition generated by the rhythmically active inhibitory interneurons. During this phase neurons are little susceptible to excitatory inputs because EPSPs are shunted and because the membrane potential is far from threshold. Hence, neurons are unlikely to respond to pre-synaptic excitatory drive. These periodic modulations of excitability can be exploited to gate communication among neurons. By adjusting oscillation frequency and phase of coupled neuronal populations communication among those neurons can either be facilitated or blocked. To form a functional network of distributed neurons it suffices to coordinate their oscillatory activity in a way that assures that signals emitted by neurons of this network are timed such that they impinge on other members of the network at times when these are highly susceptible to input. One way to achieve this is to entrain the neurons that should be bound into a functional network to engage in oscillations of the same frequency, to synchronize these oscillations and to adjust the phases such that neurons that ought to be able to communicate can communicate.

Evidence from multi-site recordings indicate that neurons are indeed bound together into sometimes widespread functional networks by synchronization of their oscillatory activity in a task dependent way (Salazar *et al.*, 2012; Buschman *et al.*, 2012). This supports the hypothesis (Gray *et al.*, 1989; Singer, 1999) that synchronization of oscillatory neuronal activity is a versatile mechanism for the temporary association of distributed neurons and the binding of their responses into functionally coherent assemblies which as a whole represent a particular cognitive content. Such a dynamic binding mechanism appears as an economical and highly flexible strategy to cope with the representation of the virtually unlimited variety of feature constellation characterizing perceptual objects. Taking the unified nature of conscious experience and the virtually infinite diversity of possible contents that can be represented, the formation of distributed representations by response synchronization offers itself as a mechanism allowing for the encod-

ing of ever changing constellations of contents in a unifying format. Synchronization is also ideally suited to contribute to the selection of contents for access to consciousness. Synchronization enhances the saliency of signals by concentrating spike discharges to a narrow temporal window. This increases the coincidence of EPSPs in target cells receiving input from synchronized cell groups. Because coincident EPSPs summate much more effectively than temporally dispersed EPSPs, synchronized inputs are particularly effective in driving post-synaptic target cells. It is thus not unexpected that entrainment of neuronal populations in synchronized gamma oscillation is used for attention dependent selection of input configurations (Fries *et al.*, 2001a; Fries, 2009).

If activation patterns that subjects can become aware of are indeed characterized by globally coherent states of those cortical regions that process the contents actually appearing as unified, one does expect that these states of awareness are associated with large scale synchronization of neuronal activity. Candidate frequency bands are the gamma and beta oscillations as these have been shown to serve the temporal coordination of cortical networks. By contrast, if subjects are not aware of the presented stimulus material, processing should remain confined to smaller subnetworks that operate in relative isolation and do not get integrated into globally coherent states. In this case one should observe only local synchronization of more circumscribed neuronal populations.

Finally, adjustments of oscillation frequency and phase fulfill the requirement that assemblies representing consciously processed contents need to be reconfigured at an extremely fast rate. The contents that subjects are aware of can change at a rapid pace, at least four times a second, if one considers that this is the frequency with which the direction of gaze changes during the scanning of natural scenes. Thus, assemblies representing contents that are consciously perceived must be reconfigurable at similarly fast time scales. Evidence suggests that cortical networks operate in a regime of self-organized criticality close to the edge of chaos (Shew *et al.*, 2009). Dynamical systems operating in this range can undergo very rapid state changes characterized by shifts in oscillation frequencies, synchronization and phase.

Methodological caveats

The most frequently used strategy for the identification of the NCC is the contrastive analysis. One creates perceptual conditions in which targets are consciously perceived only in a subset of trials while making sure that physical conditions are kept as constant as possible. This strategy implies that detection tasks are designed that operate close to perceptual threshold. This

can be achieved by reducing the physical energy of the stimuli or by masking them. While subjects are engaged in such detection tasks, neuronal responses are measured and then trials are sorted depending on whether the subjects did or did not perceive the stimulus. By subtracting the average responses obtained in the two conditions from one another, those neuronal responses can be isolated that occur only in the condition of successful detection and these are then commonly interpreted as neuronal correlate of conscious perception. This seemingly simple approach is not without ambiguity. Thus, noise fluctuations in afferent pathways are likely to lead to significant differences in the available sensory evidence, especially since experiments are performed at perceptual threshold. Therefore, those aspects of neuronal responses that truly reflect the NCC may be contaminated by signals resulting from noise fluctuations at processing stages preceding those actually mediating awareness. Also, once subjects have become aware of stimuli, there are a number of subsequent processing steps that need not necessarily be linked to the NCC. These comprise the covert verbalization of stimulus material, the engagement of working memory, the transfer of information into declarative memory and perhaps also the preparation of covert motor responses. The distinction between these various confounding factors is difficult because all these processes are intimately related to each other. A detailed discussion of this problem is given in Aru *et al.* (2012a). One distinguishing feature could be the latency of the electrographic signatures of these various processing steps. Noise dependent fluctuations in sensory evidence should be manifest early on, responses related to the NCC proper should have some intermediate latencies and the consequences of having become aware of a stimulus should have the longest latencies. In order to use these latencies as distinguishing criterion, it is of course required to estimate the precise latency at which the mechanisms leading to conscious perception are likely to be engaged. Assuming that the time required to prepare and execute simple motor responses is rather constant, the interval of interest can be constrained and has been proposed to be somewhere between 180 and a few hundred milliseconds, depending on the sensory modality and the difficulty of the detection task. Attempts to use latency criteria for the elimination of confounds is of course restricted to electroencephalographic and magneto-encephalographic data and cannot be applied to results obtained with functional magnetic resonance imaging because of the limited temporal resolution of this technique.

Another option for the reduction of confounds is to combine manipulations that influence the conscious perception of a stimulus through different mechanisms and to compare the electrographic responses between

conditions (Aru *et al.*, 2012b**)**. We applied this strategy in investigations of patients with subdurally implanted recording electrodes located over the visual cortex. In one set of trials the visibility of stimuli, in this case faces, was manipulated by changing the sensory evidence of the stimulus material. In another set of trials visibility of the same stimuli was influenced by allowing the subjects to familiarize themselves with some of the stimuli. This also facilitates detectability but now because of an expectancy driven top-down process. The reasoning was that neuronal responses reflecting the NCC proper should be the same irrespective of whether stimuli were consciously perceived because of enhanced sensory evidence or because of top-down facilitation. As electrographic signature of interest we analyzed the neuronal activity in the gamma band. In a previous study (Fisch *et al.*, 2009) had shown that category specific gamma band responses in the visual cortex correlate with conscious perception. Conscious recognition led to a phasic enhancement of the gamma band response supporting the notion that conscious perception arises locally within sensory cortices which is in line with previous conclusions (Zeki, 2001; Malach, 2007). In our study we found that the reports of the subject were clearly modulated both by changing sensory evidence and by prior knowledge of the stimuli, as expected, but the gamma band responses were solely reflecting sensory evidence. This suggests that the differential activation of specific areas of the visual cortex, in our case mainly the fusiform face area, reflect processes that prepare access to conscious perception but are not its substrate proper.

Another frequently used paradigm in the search for the NCC is interocular rivalry. If the two eyes are presented with stimuli that cannot be fused into one coherent percept, subjects perceive only one of the two stimuli at a time and these percepts alternate. There are various ways to label the stimuli presented to the two eyes, to trace the responses related to their processing in the brain and then to see which brain structures need to get involved in order to support conscious perception. Again, these studies have led to inconclusive results. Some claim that suppression of signals corresponding to the non-perceived stimulus occurs only at very high levels of visual processing as for example the temporal cortex, the highest stage of the ventral processing stream. The conclusion from these studies is that activation of this particular cortical network is a necessary prerequisite for conscious processing (Logothetis *et al.*, 1996; Silver & Logothetis, 2004). Others, by contrast, found diverging activity patterns already at the level of the thalamus and the primary visual cortex (Haynes *et al.*, 2005; Fries *et al.*, 1997). Recent correlations between the dynamics characterizing binocular rivalry and anatomical features of the primary visual cortex actually provide

compelling evidence, that the rivalry phenomenon is based on processes occurring within V1 (Genc *et al.*, 2013, under revision). However, none of these studies allows one to unambiguously locate the processes that lead to conscious perception. They only contribute to the identification of the earliest levels of processing in which changes are detectable that correlate with conscious perception.

Interesting and of potential relevance for interpretations given in the next chapter is the observation that the access of sensory signals to conscious processing does not seem to be gated by modulation of the neurons' discharge rate but rather by changes of the synchronization of their activity – at least at early stages of processing. What mattered was the degree of synchronicity of oscillatory activity in the gamma frequency range. Signals conveyed by well-synchronized neuronal assemblies had access to conscious processing while signals conveyed by similarly active but purely synchronized neurons failed to do so (Fries *et al.*, 2001b). Of interest in this context is also that stimuli access conscious perception more easily if they are attended to and that attention enhances synchronization of neuronal responses in the gamma frequency band in early visual areas. Again, however, this local increase in synchrony is likely to just enhance the saliency of the neuronal responses, facilitating their propagation across the cortical networks and cannot per se be considered as a neuronal correlate of consciousness.

Evidence relating long-range synchronization and consciousness

The results described in the following were obtained in a study where we presented words that could be perceived in some trials and not in others (by adjusting the luminance of masking stimuli) and simultaneously performed electroencephalographic (EEG) recordings (Melloni *et al.*, 2007). Several measures were analyzed: Time-resolved power changes of local signals, the precision of phase synchronization across recording sites over a wide frequency range, and event-related potentials (ERPs). A brief burst of long distance synchronization in the gamma frequency range between occipital, parietal and frontal sensors was the first event that distinguished seen from unseen words at about 180ms poststimulus. In contrast local synchronization was similar between conditions. Interestingly, after this transient period of synchronization, several other measures differed between seen and unseen words: We observed an increase in amplitude of the P300 ERP for visible words which most likely corresponds to the transfer of information to working memory. In addition, during the interval period in which visible words had to be maintained in memory, we observed increases in frontal theta oscillations. Theta oscillations have been related to maintenance

of items in short-term memory (Jensen & Tesche, 2002; Schack *et al.*, 2005).

To test whether the increase in long-distance synchronization relates to awareness or depth of processing, we further manipulated the depth of processing of invisible words. It has previously been shown that invisible words can be processed up to the semantic and motor level (Dehaene *et al.*, 1998). In a subliminal semantic priming experiment we briefly presented words (invisible) that could either be semantically related or not related with a second visible word on which subjects had to carry out a semantic classification task. Invisible words were processed up to semantic levels as revealed by modulation of the reaction times depending on the congruency between invisible and visible words: Congruent pairs exhibited shorter reaction times than incongruent ones. We observed increases in power in the gamma frequency range for unseen but processed words. For visible words we additionally observed increases in long-distance synchronization in the gamma frequency range (Melloni & Rodriguez, 2007). Thus, local processing of stimuli is reflected in increases in gamma power, whereas long-distance synchronization seems to be related to awareness of the stimuli. This suggests that conscious processing requires a particular dynamical state of the cortical network. The large-scale synchronization that we observed in our study could reflect the transfer of contents into awareness and/or their maintenance. We favor the first possibility given the transient nature of the effect and argue that the subsequent theta oscillations might support maintenance. It is conceivable that short periods of long-distance synchronization in the gamma band reflect the update of new contents, while the slower pace of theta oscillations might relate to sustained integration and maintenance of local results in the workspace of consciousness. The interplay between these two frequency bands might underlie the phenomenon of continuous but ever changing conscious experience (see below).

More recently, Gaillard *et al.* (2009) revisited the question of processing of visible and invisible words. In intracranial recordings in epileptic patients they observed that invisible words elicited activity in multiple cortical areas which quickly vanished after 300 ms. In contrast, visible words elicited sustained voltage changes, increases in power in the gamma band, as well as long-distance synchronization in the beta band and long-range Granger causality. In contrast to our study, Gaillard *et al.* observed a rather late (300-500 ms) rise of long-distance synchronization. However, it is important to note that in the study of Gaillard *et al.*, phase-synchrony was analyzed mostly over electrodes within a given cortical area or at most between hemispheres. It is thus conceivable that earlier synchronization events passed undetected because of incomplete electrode coverage. Despite these restrictions, this

study provides one of the most compelling pieces of evidence for a relation between long-distance synchronization and consciousness.

Some results of the experiments on binocular rivalry point in the same direction. Several studies have shown increased synchronization and phase locking of oscillatory responses to the stimulus that was consciously perceived and controlled behavior (Cosmelli *et al.*, 2004; Doesburg *et al.*, 2005; Fries *et al.*, 1997; Srinivasan *et al.*, 1999). Cosmelli *et al.* (2004) extended the findings obtained in human subjects by performing source reconstruction and analyzing phase-synchrony in source space. These authors observed that perceptual dominance was accompanied by co-activation of occipital and frontal regions, including anterior cingulate and medial frontal areas. Recently, Doesburg *et al.* (2009) provided evidence for a relation between perceptual switches in binocular rivalry and theta and gamma band synchronization. Perceptual switches were related to increments in long-distance synchronization in the gamma band between several cortical areas (frontal and parietal) that repeated at the rate of theta oscillations. The authors suggested that transient gamma-band synchronization supports discrete moments of perceptual experience while theta oscillations structure their succession in time, pacing the formation and dissolution of distributed neuronal assemblies. Thus, long-range gamma synchronization locked to ongoing theta oscillations could serve to structure the flow of conscious experience allowing for changes in content every few hundred millisecond. Further research is required to clarify the exact relation between the two frequency bands and their respective role in the generation of percepts and the pacing of changes in perception.

Another paradigm in consciousness research exploits the attentional blink phenomenon. When two stimuli are presented at short intervals among a set of distractors, subjects usually detect the first (S1) but miss the second (S2) when it is presented 200-500 ms after S1. Increases in long-range neuronal synchrony in the beta and gamma frequency ranges have been observed when the S2 is successfully detected (Gross *et al.*, 2004; Nakatani *et al.*, 2005). Furthermore, Gross *et al.* (2004) observed that successful detection of both S1 and S2 was related to increased long-distance synchronization in the beta range to both stimuli, and this enhanced synchrony was accompanied by higher de-synchronization in the inter-stimulus-interval. Thus, de-synchronization might have facilitated the segregation of the two targets, allowing for identification of the second stimulus (also see Rodriguez *et al.*, 1999). Source analysis revealed, as in the case of binocular rivalry, dynamical coordination between frontal, parietal, and temporal regions for detected targets (Gross *et al.*, 2004).

In summary, studies of masking, binocular rivalry, and the attentional blink support the involvement of long-range synchronization in conscious perception. Recent investigations have suggested further that a nesting of different frequencies, in particular of theta and gamma oscillations, could play a role in pacing the flow of consciousness. Furthermore, the study of Gross *et al*. (2004) suggests that de-synchronization could serve to segregate representations when stimuli follow at short intervals. These results are encouraging and should motivate further search for relations between oscillatory activity in different frequency bands and consciousness, whereby attention should be focused not only on the formation of dynamically configured networks but also on their dissolution.

Conclusions and outlook

Large-scale synchronization of oscillatory activity has been identified as one candidate for the NCC. This variable has the advantage that it can be measured relatively directly in humans, who are able to give detailed descriptions about their conscious experience. However, oscillations and synchrony seem to be mechanisms that are as intimately and inseparably related to neuronal processing in general just as the modulation of neuronal discharge rates. Thus, without further specification these phenomena cannot stand up as NCC apart from the triviality that consciousness does not exist without them. We propose that the spatial scale and the precision and stability of neuronal synchrony might be taken as more specific indicators of whether the communication of information in the brain is accompanied by conscious experience or not. In this framework, conscious experience arises only if information that is widely distributed within or across subsystems is not only processed and passed on to executive structures but bound together into a coherent, all-encompassing, non-local but distributed meta-representation. This interpretation is compatible with views considering consciousness as the result of the dynamic interplay of brain subsystems that allows for a rapid and highly flexible integration of information provided by the numerous distributed subsystems that operate in parallel. This view resembles the proposal of Sherrington formulated in his book *The Integrative Action of the Nervous System* (Sherrington, 1904): "Pure conjunction in time without necessarily cerebral conjunction in space lies at the root of the solution of the problem of the unity of mind". The additive value of conscious processing would then be the possibility to establish in a unified data format ever changing relations between cognitive contents irrespective of whether they are read out from memory or induced by sensory signals. By virtue of this dynamic definition of novel relations, non-local meta-representations

of specific constellations could be established that have the status of cognitive objects. Just as any other distributed representation of contents these could then be stored as distributed engrams by use dependent modification of synaptic connections and influence future behavior. Thus, conscious processing would differ from non-conscious processing because it allows for the versatile binding of the previously unbound into higher order representations. If so, "conscious" processing would be functionally relevant and not merely an epiphenomenon.

References

Aru, J., Bachmann, T., Singer, W., & Melloni, L. (2012a). Distilling the neural correlates of consciousness. *Neurosci Biobehav Rev*, *36*, 737-746.

Aru, J., Axmacher, Do Lam, A.T.A., Fell, J., Elger, C.E., Singer, W., & Melloni, L. (2012b). Local category-specific gamma band responses in the visual cortex do not reflect conscious perception. *J Neurosci*, *32*(43), 14909-14914.

Baars, B.J., (1997). *In the theatre of consciousness.* Global workspace theory, a rigorous scientific theory of consciousness. *jcs*, *4*(4), 292-309.

Buschman, T.J., Denovellis, E.L., Diogo, C., Bullock, D., & Miller, E.K. (2012). Synchronous oscillatory neuronal ensembles for rules in the prefrontal cortex. *Neuron, (76)*4, 838-846.

Chalmers, D.J. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.), *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (pp. 17-40). Cambridge, MA, MIT Press.

Cosmelli, D., David, O., Lachaux, J.P., Martinerie, J., Garnero, L., Renault, B., *et al.* (2004). Waves of consciousness: ongoing cortical patterns during binocular rivalry. *Neuroimage*, *23*(1), 128-140.

Cowey, A., & Stoerig, P. (1991). The neurobiology of blindsight. *Trends Neurosci, 14*, 140.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Semin Neurosci*, 2, 263-275.

Dehaene, S., Changeux, J.P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn Sci*, *10*(5), 204-211.

Dehaene, S., Naccache, L., Le Clec, H.G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., *et al.* (1998). Imaging unconscious semantic priming. *Nature*, *395*(6702), 597-600.

Dennet, D.C. (1992). *Consciousness Explained.* London, Penguin.

Doesburg, S.M., Green, J.J., McDonald, J.J., & Ward, L.M. (2009). Rhythms of consciousness: binocular rivalry reveals large-scale oscillatory network dynamics mediating visual perception. *PLoS One*, *4*(7), e6142.

Doesburg, S.M., Kitajo, K., & Ward, L.M. (2005). Increased gamma-band synchrony precedes switching of conscious perceptual objects in binocular rivalry. *Neuroreport, 16*(11), 1139-1142.

Fisch, L., Privman, E., Ramot, M., Harel, M., Nir, Y., Kipervasser, S., Andelman, F., Neufeld, M.Y, Kramer, U., Fried, I., Malach, R. (2009). Neural "ignition": enhanced activation linked to perceptual awareness in human ventral stem visual cortex. *Neuron*, *64*, 562-574.

Fries, P. (2009). Neuronal gamma-band syn-

chronization as a fundamental process in cortical computation. *Annu Rev Neurosci*, *32*, 209-224.

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn* Sci, *9*(10), 474-480.

Fries, P., Reynolds, J.H., Rorie, A.E., & Desimone, R. (2001a). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, *291*, 1560-1563.

Fries, P., Schröder, J.H., Singer, W., & Engel, A.K. (2001b). Conditions of perceptual selection and suppression during interocular rivalry in strabismic and normal cats. *Vis Res* 41, 771-783.

Fries, P., Roelfsema, P.R., Engel, A.K., König, P., & Singer, W. (1997). Synchronisation of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proc Natl Acad Sci U S A, 101*(35), 13050-13055.

Gaillard, R., Dehaene, S., Adam, C., Clemenceau, S., Hasboun, D., Baulac, M., *et al.* (2009). Converging intracranial markers of conscious access. *PLoS Biol*, 7(3), e61.

Genc, E., Bergmann, J., Singer, W., & Kohler, A. (2013). Surface area of early visual cortex predicts individual speed of travelling waves during binocular rivalry. *J Neurosci.*, under revision.

Gray, C.M., König, P., Engel, A.K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, *338*, 334-337.

Gray, C.M., & Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc Natl Acad Sci U S A, 86*, 1698-1702.

Gross, J., Schmitz, F., Schnitzler, I. Kessler, K., Shaprio, K., Hommel, B., *et al.* (2004). Modulation of long-range neural synchrony reflects temporal limitations of visual attention in humans. *Proc Natl Acad Sci U S A, 101*(35), 13050-13055.

Hameroff, S. (2006). "Consciousness, neurobiology and quantum mechanics". In Tuszynski, Jack, *The Emerging Physics of Consciousness* (pp. 193-253). Berlin, Springer.

Han, X., Chow, B.Y., Zhou, H., Klapoetke, N.C., Chuong, A.S., Rajimehr, R., Yang, A., Baratta, M.V., Winkle, J., Desimone, R., & Boyden, E.S. (2011). A high-light sensitivity optical neural silencer: development and application to optogenetic control of non-human primate cortex. *Frontiers Syst Neurosci*, *5*(18), 1-8.

Haynes, J.-D., Deichmann, R., & Rees, G. (2005). Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus, *Nature,* 438*, 496-499.*

Houweling, A.R. & Brecht, M. (2008). Behavioural report of single neuron stimulation in somatosensory cortex. *Nature*, *451*, 65-68.

Jensen, O., & Tesche, C.D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *Eur J Neurosci, 15*(8), 1395-1399.

Lau, H.C., & Passingham, R.E. (2007). Unconscious activation of the cognitive control system in the human prefrontal cortex. *J Neurosci*, *27*(21), 5805-5811.

Logothetis, N.K., Leopold, D.A., & Sheinberg, D.L. (1996). What is rivalling during binocular rivalry? *Nature, 380*, 621-624.

Malach, R. (2007). The measurement problem in human consciousness research. *Behav Brain Sci*, *30*, 516-517.

Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W., & Rodriguez, E. (2007). Synchronization of neural activity across cortical areas correlates with conscious perception. *J Neurosci*, 27*(11), 2858-2865.*

Melloni, L., & Rodriguez, E. (2007). Non-perceived stimuli elicit lobal but not large-scale neural synchrony. *Perception*, *36* (ECVP Abstract Supplement).

Metzinger, T. (2000). *Neural Correlates of Consciousness: Empirical and Conceptual*

Questions. Cambridge, MA, MIT Press.

Nakatani, C., Ito, J., Nikolaev, A.R., Gong, P., & van Leeuwen, C. (2005). Phase synchronization analysis of EEG during attentional blink. *J Cogn Neurosci, 17*(12), 1969-1979.

Penrose, R. (1994). *Shadows of the mind: A search for the missing science of consciousness*. USA, Oxford University Press.

Rodriguez, E., George, N., Lachaux, J.P., Martinerie, J., Renault, B., & Varela, F.J. (1999). Perception's shadow: long distance synchronization of human brain activity. *Nature*, *397*(6718), 430-433.

Salazar, R.F., Dotson, N.M., Bressler, S.L., & Gray, C.M. (2012). Content specific fronto-parietal synchronization during visual working memory. *Science, 338*, 1097-1100.

Salzman, C.D., Murasugi, C.M., Britten, K.H., & Newsome, W.T. (1992). Microstimulation in visual area MT: Effects on direction discrimination performance. *J Neurosci*, *12*, 2331-2355.

Schack, B., Klimesch, W., & Sauseng, P. (2005). Phase synchronisation between theta and upper alpha oscillations in a working memory task. *Int J Psychophysiol*, *57*(2), 105-114.

Searle, J.R. (1997). *The Mystery of Consciousness.* London, Granta Books.

Sherrington, C.S. (1906). *The Integrative Action of the Nervous System*. New York, Charles Scribner's Sons.

Shew, W.L., Yang, H., Petermann, T., Roy, R., & Plenz, D. (2009). Neuronal avalanches imply maximum dynamic range in cortical networks at criticality. *J Neurosci*, *29*, 15595-15600.

Silver, M.A.; & Logothetis, N.K. (2004). Grouping and segmentation in binocular rivalry. *Vis Res, 44*, 1675-1692.

Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron, 24*, 49-65.

Srinivasan, R., Russell, D.P., Edelman, G.M., & Tononi, G., (1999). Increased synchronization of neuromagnetic responses during conscious perception. *J Neurosci, 19*(13), 5435-5448.

Uhlhaas, P.J., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolic, D., & Singer, W. (2009). Neuronal synchrony in cortical networks: history, concept and current status. *Frontiers Integrat Neurosci*, *3*, 1-19.

Van den Heuvel, M.P., & Sporns, O. (2011). Rich club organization of the human connectome. *J Neurosci*, *31*, 15775-15786.

Van Gaal, S., Ridderinkhof, K.R., Fahrenfort, J.J., Scholte, H.S., & Lamme, V.A. (2008). Frontal cortex mediates unconsciously triggered inhibitory control. *J Neurosci*, *28*(32), 8053-8062.

Varela, F. Lachaux, J.P., Rodriguez, E., & Martinerie, J. (2001). The brainweb: phase synchronisation and large-scale integration. *Nat Rev Neurosci*, *2*(4), 229-239.

Von der Malsburg, C., Phillips W.A., & Singer, W. (2010). *Dynamic Coordination in the Brain. From Neurons to Mind*. Cambridge, MA & Frankfurt a.M., MIT Press & FIAS.

Zeki, S. (2001). Localization and globalization in conscious vision. *Annu Rev Neurosci, 24*, 57-86.

Brain Rhythms for Cognition and Consciousness

Earl K. Miller[1] and Timothy J. Buschman[2]

What does a thought look like? In neurobiological terms, it is universally assumed to involve a neural ensemble – a subset of neurons that together represent an item of information, whether it is a computation, memory, percept, or desire. This is a bit of a truism. If neurons represent information, then a subset of them must represent a given thought. What is not so obvious is how the brain forms ensembles. Neurons, the brain's basic processing units, do not have single functions; they do not contribute to just one thought or memory. Rather, neurons, especially at the higher areas of cortex central to cognition, are highly multivariate and dynamic. Neurons "multitask": They process a wide range of often seemingly unrelated information that can contribute to many different functions and computations. In other words, neurons do not participate in a single ensemble. They participate in many overlapping ensembles. Further, consider that intelligent, goal-directed thought and action requires integration of a wide range of information, not only about our external environment but also our internal state, relevant stored knowledge, possible courses of action, and anticipated outcomes. It thus seems unlikely that normal, rational thought and action stems from neural activity haphazardly bouncing around the brain's many networks. The brain must have mechanisms that coordinate interactions among its neurons in order to form the ensembles and networks of ensembles that produce clear and coherent thought and action.

Here, we discuss evidence that the brain regulates the flow of neural "traffic" via rhythmic synchrony between neurons. Neurons form ensembles, and ensembles become part of larger functional networks, when they "hum" together. Conversely, they don't form ensembles, and don't interact, when they don't hum together. In other words, rhythmic synchrony can reinforce *or* prevent communication between neurons. Synchronizing the rhythmic activity of two sets of neurons ensures they are in the excited state at the same time

[1]The Picower Institute for Learning and Memory and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA.

[2]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ, USA.

and therefore primed to transmit information to each other. It follows that desynchronization of these rhythms would actively interfere with communication. Thus, by changing the rhythmic synchronization between neurons, their communication can be altered, changing the flow of information through the brain. Synchronized brain rhythms may also explain the most obvious and objective fact about consciousness: it is very hard, often impossible, to think about more than one or a few things at the same time.

Synchronized rhythms can regulate network interactions

It has long been known that the brain has large populations of neurons that oscillate in synchrony. These so-called "brain waves" occur across a wide range of frequencies from very low (< 1Hz) to very high (>60 Hz). They have long been known to vary with mental state. More relaxed states produce lower frequency waves and increased cognitive effort produces higher frequency waves. But, for many years, their exact role in brain function has been a mystery. This is largely because much of our understanding of brain networks has been inferred largely from indirect evidence such as anatomical connections and properties of the brain's individual parts studied in isolation. This modular understanding stands to reason. Identifying and characterizing the brain's components is prerequisite to any integrated understanding of the whole. And technological limitations had largely restricted us to piecemeal investigation. But technical and methodological advancements have led to increasing investigation and understanding at the network level. Beginning with the pioneering work of von der Malsburg, Singer and colleagues, there has been increasing awareness that the precise synchrony of timing of activity *between* neurons may be critical in forming functional networks.

Oscillations are useful for producing synchrony between neural impulses ("spikes"). Spikes from two neurons that arrive simultaneously at a third, downstream, neuron have a bigger impact than if the impulses arrived at different times (Aertsen *et al*. 1989; Salinas and Sejnowski 2000; Pascal Fries 2005; Engel, Fries, and Singer 2001). Thus, synchronicity between neurons can improve signal to noise ratio of neural signals while, at the same time, reducing the number of spikes (spikes cost energy) needed to represent a stimulus (Tiesinga *et al*. 2002; Siegel and König 2003; Aertsen *et al*. 1989; Azouz and Gray 2000). If true, then one might expect the activity of synchronized neurons to carry more information than non-synchronized neurons. This has been seen in the prefrontal cortex (Siegel, Warden, and Miller 2009) and, in fact, only synchronous neurons in parietal cortex predict behavior in a reach-and-saccade task (Dean, Hagan, and Pesaran 2012).

Evidence of the role of rhythmic synchrony in boosting neural signals comes from studies of visual attention. Increased attentional focus is associated with changes in oscillatory synchrony in sensory cortex. Visual cortical neurons that process a stimulus under attentional focus show increased synchronized gamma band (30-90 Hz) oscillations (P. Fries *et al.* 2001). By contrast, neurons representing an unattended stimulus showed increased low frequency (<17 Hz) synchronization. A variety of evidence suggests that low frequencies may help deselect or inhibit the corresponding ensembles (Buschman *et al.* 2012; Vijayan and Kopell 2012; Palva and Palva 2011; Ray and Cole 1985). Higher frequency (>30 Hz) synchrony may result from local interactions within a cortical area (Cardin *et al.* 2009; Börgers, Epstein, and Kopell 2008), the same interactions that underlie the computations of stimulus features within a cortical area (Wilson *et al.* 2012; Lee *et al.* 2012; Reynolds and Heeger 2009). The idea is that attention boosts the high frequency synchrony of neurons processing an attended object, thus boosting its impact on other neural processing. Supporting this hypothesis, microstimulation of the frontal eye fields induces high-frequency oscillations in parietal cortex neurons processing an attended stimulus (Premereur *et al.* 2012).

Synchrony between regions may also regulate communication across large-scale brain-wide network interactions. If two different networks in different brain areas oscillate in phase they are more likely to influence one another because, as noted above, they are both in an excited and receptive state at the same time. Conversely, if they are out of phase, they are less likely to influence each other. This has led to the suggestion that oscillatory synchrony could be used to regulate communication between brain areas (Bressler 1996; Engel, Fries, and Singer 2001; Salinas and Sejnowski 2000; Pascal Fries 2005). Support for this notion comes from observations that inter-areal oscillatory coherence within and between "cognitive" regions and sensory areas has been found to increase with attention (Buschman and Miller 2007; Saalmann, Pigarev, and Vidyasagar 2007; Siegel *et al.* 2008; Gregoriou *et al.* 2009).

Synchronized rhythms form neural ensembles

Above, we discussed how synchronized rhythms can boost neural signals and regulate which networks in the brain "talk" to one another. The same mechanisms can also play a role in forming neural ensembles. Just as humming together allows networks across the brain to communicate more effectively, individual neurons may form local networks of ensembles when their synchronized humming reinforces their mutual communication.

Some sort of mechanism for selecting specific neurons for membership in specific ensembles must exist. As mentioned above, many neurons in higher cortex are very heterogeneous and many seem to "multiplex". They signal all sorts of seemingly unrelated information at different times. For example, at one time a given neuron may convey information about the concept "cat" whereas a few moments later, it might represent an entirely different category or precept or even seem to be issuing motor commands (Rainer, Rao, and Miller 1999; Cromer *et al.* 2011). Computational modeling suggests that this mixed selectivity allows the system to encode a large variety of memories, memories, events, rules, etc. with a biologically realistically limited number of neurons (Rigotti *et al.* 2010).

But this diversity seems to work against the demand to activate a specific ensemble that represents a specific thought. If higher cortical neurons have many connections reflecting a wide range of information, why doesn't neural activity simply run around the connections and activate many ensembles in a jumble? Synchrony between neighboring neurons can dynamically "carve" an ensemble from a greater, heterogeneous, population of neurons (Akam and Kullmann 2010) by reinforcing mutual activation between the neurons that form the ensemble (Womelsdorf *et al.* 2007). Because ensemble membership would depend on which neurons are oscillating in synchrony at a given moment, ensembles could flexibly form, break-apart, and re-form *without changing the physical structure* of the underlying neural network. In other words, this may endow ensembles with a critical feature: flexibility in their construction. Flexibility is a hallmark of higher cognition. Humans can quickly adapt and change their thoughts and behaviors in order to tailor them to the constantly changing demands of our complex world. Thus, ensembles have to be assembled, deconstructed, and reconfigured from moment to moment. Synchronized oscillations can provide the substrate.

To test this, we recently examined neural activity in the prefrontal cortex (PFC) of monkeys switching between two cognitively-demanding tasks (Buschman *et al.* 2012). The PFC is critical for cognitive flexibility. When it is damaged or dysfunctional, people are unable to suppress prepotent, reflexive reactions in favor of a more contextually-appropriate response (Owen *et al.* 1993; Bechara, Tranel, and Damasio 2000). Furthermore, patients with PFC damage often perseverate, inappropriately repeating a particular behavior or line of thought (Barceló and Knight 2002; Rossi *et al.* 2007; Milner 1963).

Monkeys were trained to switch between paying attention to either the color (either red or blue), or orientation (either horizontal or vertical) of a line. We measured fluctuations in local field potentials (LFPs) at different

points along the prefrontal cortex, from an array of electrodes spaced 1mm apart. LFPs are the summed activity of many neurons near the recording electrode, like the brain waves that can be recorded from the human scalp. When the monkeys focused their attention on the task, there was an increase in oscillations at high frequencies in so-called beta waves (19-40 Hz). Depending on which rule was in effect (i.e. whether the monkeys were paying attention to either color or orientation) different patterns of electrodes were synchronized at these beta waves. Some neuron clusters overlapped, belonging to more than one group, but each pattern of beta-wave synchrony had its own distinctive pattern. In other words, beta wave synchrony seemed to establish different, but physically overlapping ensembles across the prefrontal cortex.

We also observed oscillations in the low-frequency alpha range (6-16 Hz) among neurons that formed the orientation rule ensemble. However, this only happened when the monkey was preparing to apply the other rule, color. Alpha waves have been associated with suppression or inattention (Haegens *et al.* 2011; Gould, Rushworth, and Nobre 2011) and thus may create an inhibition of irrelevant processes (Klimesch, Sauseng, and Hanslmayr 2007; Mathewson *et al.* 2011). In our case, these alpha oscillations seemed to be acting to quiet the neurons that formed the orientation rule ensemble when the animal was preparing to do the opposite, color, rule. This alpha suppression was necessary because orientation seemed to be the dominant modality for the monkeys. Whenever they switched from paying attention to orientation to color, they cognitively "stumbled"; that is, their behavioral reactions slowed temporarily. By contrast, there was no stumbling when they switched from color to orientation. This suggests that orientation had a naturally greater hold on the animals' attention than color. This may be due to its relative saliency, much like word-naming in the Stroop test (MacLeod 1991).

This all suggests that synchronous oscillations helped control the formation of ensembles (Kopell, Whittington, and Kramer 2011). Higher (beta) frequencies defined the two rule ensembles (pay attention to color vs orientation) while lower (alpha) frequencies were used to somehow disrupt formation of the stronger ensemble (and thus prevent an erroneous reflexive reaction) when the weaker ensemble had to be used.

If synchronized rhythms form neural ensembles, one might naturally wonder how it is that the brain can form more than one ensemble at a time. After all, would not two rhythmically defined ensembles inadvertently synchronize to each other, merging together and distorting the information each other represents? In fact, the brain does have a great deal of trouble

having more than one or (at most) a few ensembles simultaneously activated in consciousness. Humans have a very small capacity for simultaneous thoughts; it is a defining feature of consciousness (and the reason why one should not drive and use a mobile phone at the same time). As we will see next, this may be because the brain keeps ensembles from interfering with one another by having them oscillate out of phase with one another. In other words, consciousness may be a mental juggling act, and only a few balls can be juggled at once.

Synchronized rhythms, capacity limits, and consciousness

The finite resources of cognition have been well-known since the classic George Miller paper describing the capacity of working memory as the "magic number" of seven plus or minus two (Miller 1956). More recent work has lowered the magic number to four or five for the average adult human (Cowan 2001). The exact capacity of a person varies from individual to individual; some can remember only 1-2 items and others can remember up to 7 (Vogel and Machizawa 2004; Vogel, McCollough, and Machizawa 2005). An individual's capacity is highly correlated with measures of fluid intelligence, reflecting the fact that these capacity limits are a fundamental restriction in high-level cognition (Fukuda *et al.* 2010; Conway, Kane, and Engle 2003). This makes sense: the more thoughts that can be simultaneously held "in mind" and manipulated, the more associations, connections, and relationships can be made, and the more sophisticated thought can be.

Thus, there seems to be a fundamental limitation in the number of separate items that can be represented simultaneously in neural activity, particularly in an active state that is accessible to high-level cognition. An explanation readily follows when neural ensembles are formed via synchronized rhythms. The idea is that neurons that are part of a specific ensemble tend to align their spikes to specific phases of neuronal population oscillations (O'Keefe and Recce 1993; Hopfield and Herz 1995; Laurent 2002; Mehta, Lee, and Wilson 2002; König and Engel 1995; P. Fries, Nikolic, and Singer 2007).

Multiple items are simultaneously held "in mind" by multiplexing them at different phases of population oscillations (Figure 1) (Jensen and Lisman 2005; Lisman and Idiart 1995). In other words, the mechanisms for conscious thought "juggle" separate ensembles by oscillating them out of phase of one another. Evidence for this multiplexing when information is held "in mind" was recently reported by Siegel *et al.* (Siegel, Warden, and Miller 2009). When monkeys held multiple objects in working memory, prefrontal neurons encode information about each object at different phases of an ongoing, ~32 Hz, oscillation. This phase-based coding has an inherent capacity

limitation because, presumably, only so much information can fit within an oscillatory cycle (that is, only a few "balls" can be juggled at once). Crucial tests of this hypothesis still need to be conducted, but all this suggests that making thoughts conscious may depend on generation of oscillatory rhythms and the precise temporal relationships between them and the spiking of individual neurons.

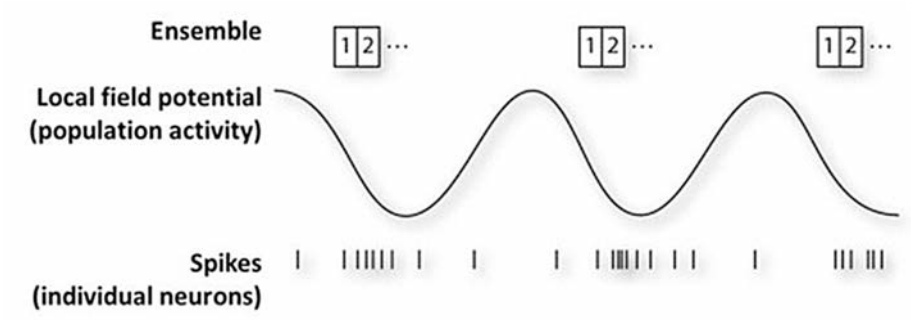

Figure 1. This figure illustrates oscillatory phase-coding. The idea is that neural ensembles of the under two simultaneous thoughts ("1" and "2"), oscillate at similar frequencies but at different phases of the oscillation. In other words, the ensembles line up on different parts of the brain wave. This may explain the severely limited capacity of consciousness. Only a few thoughts can fit in each wave.

Conclusions

We have discussed how rhythmic synchrony can provide a fundamental backbone for forming and coordinating interactions within and across disparate neural networks. The act of putting neural activity from specific neurons in precise lockstep with specific other neurons may form the ensemble representing a specific thought and regulate which ensembles and larger networks "talk" to one another. The implication is that the signals that underlie cognition do not operate continuously, but rather discretely, with pulses of activity routing packets of information. The advantage is that it constrains and shapes the flow of neural signals. In other words, the brain's physical infrastructure (i.e., its anatomy) dictates where neural signals *can* flow; synchronized rhythms dictate where signals *do* flow. However, this comes at a cost. Any coding scheme based on repeated rhythmic activity is naturally limited in bandwidth; only so many things can be computed or carried in a single oscillatory cycle. This may explain the most fundamental property of conscious thought, its limited capacity.

Acknowledgements

We thank the National Institute of Mental Health, National Science Foundation, and the Picower Foundation for support and Marlene Wicherski for comments on the manuscript.

References

Aertsen, A.M., G.L. Gerstein, M.K. Habib, and G. Palm. 1989. "Dynamics of Neuronal Firing Correlation: Modulation of 'Effective Connectivity'." *Journal of Neurophysiology* 61 (5): 900-917.

Akam, Thomas, and Dimitri M. Kullmann. 2010. "Oscillations and Filtering Networks Support Flexible Routing of Information". *Neuron* 67 (2) (July): 308-320. doi:10.1016/j.neuron.2010.06.019

Azouz, Rony, and Charles M. Gray. 2000. "Dynamic Spike Threshold Reveals a Mechanism for Synaptic Coincidence Detection in Cortical Neurons in Vivo". *Proceedings of the National Academy of Sciences* 97 (14) (July 5): 8110-8115. doi:10.1073/pnas.130200797

Barceló, Francisco, and Robert T. Knight. 2002. "Both Random and Perseverative Errors Underlie WCST Deficits in Prefrontal Patients". *Neuropsychologia* 40 (3): 349-356. doi:10.1016/S0028-3932(01)00110-5

Bechara, Antoine, Daniel Tranel, and Hanna Damasio. 2000. "Characterization of the Decision-making Deficit of Patients with Ventromedial Prefrontal Cortex Lesions". *Brain* 123 (11) (November 1): 2189-2202. doi:10.1093/brain/123.11.2189

Börgers, Christoph, Steven Epstein, and Nancy J. Kopell. 2008. "Gamma Oscillations Mediate Stimulus Competition and Attentional Selection in a Cortical Network Model". *Proceedings of the National Academy of Sciences* 105 (46) (November 18): 18023-18028. doi:10.1073/pnas.0809511105

Bressler, Steven L. 1996. "Interareal Synchronization in the Visual Cortex". *Behavioural Brain Research* 76 (1-2) (April): 37-49. doi:10.1016/0166-4328(95)00187-5

Buschman, Timothy J., Eric L. Denovellis, Cinira Diogo, Daniel Bullock, and Earl K. Miller. 2012. "Synchronous Oscillatory Neural Ensembles for Rules in the Prefrontal Cortex". *Neuron* 76 (4) (November 21): 838-846. doi:10.1016/j.neuron.2012.09.029

Buschman, Timothy J., and Earl K. Miller. 2007. "Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices". *Science* 315 (5820) (March 30): 1860-1862. doi:10.1126/science.1138071

Cardin, Jessica A., Marie Carlen, Konstantinos Meletis, Ulf Knoblich, Feng Zhang, Karl Deisseroth, Li-Huei Tsai, and Christopher I. Moore. 2009. "Driving Fast-spiking Cells Induces Gamma Rhythm and Controls Sensory Responses". *Nature* 459 (7247) (June 4): 663-667. doi:10.1038/nature08002

Conway, Andrew R.A., Michael J. Kane, and Randall W. Engle. 2003. "Working Memory Capacity and Its Relation to General Intelligence". *Trends in Cognitive Sciences* 7 (12) (December): 547–552. doi:10.1016/j.tics.2003.10.005

Cowan, Nelson. 2001. "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity". *Behavioral and Brain Sciences* 24 (01): 87-114. doi:null

Cromer, Jason A., Jefferson E. Roy, Timothy J. Buschman, and Earl K. Miller. 2011. "Comparison of Primate Prefrontal and Premotor Cortex Neuronal Activity During Visual Categorization". *Journal of Cog-*

nitive Neuroscience 23 (11) (March 31): 3355-3365. doi:10.1162/jocn_a_00032

Dean, Heather L., Maureen A. Hagan, and Bijan Pesaran. 2012. "Only Coherent Spiking in Posterior Parietal Cortex Coordinates Looking and Reaching". *Neuron* 73 (4) (February 23): 829-841. doi:10.1016/j.neuron.2011.12.035

Engel, Andreas K., Pascal Fries, and Wolf Singer. 2001. "Dynamic Predictions: Oscillations and Synchrony in Top-down Processing". *Nat Rev Neurosci* 2 (10) (October): 704-716. doi:10.1038/35094565

Fries, P., D. Nikolic, and W. Singer. 2007. "The Gamma Cycle". *Trends in Neurosciences* 30 (7) (July): 309-316. doi:10.1016/j.tins.2007.05.005

Fries, P., J.H. Reynolds, A.E. Rorie, and R. Desimone. 2001. "Modulation of Oscillatory Neuronal Synchronization by Selective Visual Attention". *Science* 291 (5508): 1560.

Fries, Pascal. 2005. "A Mechanism for Cognitive Dynamics: Neuronal Communication Through Neuronal Coherence". *Trends in Cognitive Sciences* 9 (October): 474-480. doi:10.1016/j.tics.2005.08.011

Fukuda, Keisuke, Edward Vogel, Ulrich Mayr, and Edward Awh. 2010. "Quantity, Not Quality: The Relationship Between Fluid Intelligence and Working Memory Capacity". *Psychonomic Bulletin & Review* 17 (5) (October): 673-679. doi:10.3758/17.5.673

Gould, Ian C., Matthew F. Rushworth, and Anna C. Nobre. 2011. "Indexing the Graded Allocation of Visuospatial Attention Using Anticipatory Alpha Oscillations". *Journal of Neurophysiology* 105 (3) (March 1): 1318-1326. doi:10.1152/jn.00653.2010

Gregoriou, Georgia G., Stephen J. Gotts, Huihui Zhou, and Robert Desimone. 2009. "High-Frequency, Long-Range Coupling Between Prefrontal and Visual Cortex During Attention". *Science* 324 (5931) (May 29): 1207-1210. doi: 10.1126/science.1171402

Haegens, Saskia, Verónica Nácher, Rogelio Luna, Ranulfo Romo, and Ole Jensen. 2011." -Oscillations in the Monkey Sensorimotor Network Influence Discrimination Performance by Rhythmical Inhibition of Neuronal Spiking". *Proceedings of the National Academy of Sciences* 108 (48) (November 29): 19377-19382. doi:10.1073/pnas.1117190108

Hopfield, J.J., and A.V. Herz. 1995. "Rapid Local Synchronization of Action Potentials: Toward Computation with Coupled Integrate-and-fire Neurons". *Proceedings of the National Academy of Sciences* 92 (15) (July 18): 6655-6662.

Jensen, Ole, and John E. Lisman. 2005. "Hippocampal Sequence-encoding Driven by a Cortical Multi-item Working Memory Buffer". *Trends in Neurosciences* 28 (2) (February): 67-72. doi:10.1016/j.tins.2004.12.001

Klimesch, Wolfgang, Paul Sauseng, and Simon Hanslmayr. 2007. "EEG Alpha Oscillations: The Inhibition-timing Hypothesis". *Brain Research Reviews* 53 (1) (January): 63-88. doi:16/j.brainresrev.2006.06.003

König, Peter, and Andreas K. Engel. 1995. "Correlated Firing in Sensory-motor Systems". *Current Opinion in Neurobiology* 5 (4) (August): 511-519. doi:10.1016/0959-4388(95)80013-1

Kopell, N., M.A. Whittington, and M.A. Kramer. 2011. "Neuronal Assembly Dynamics in the Beta1 Frequency Range Permits Short-Term Memory". *Proceedings of the National Academy of Sciences* 108 (9) (March 1): 3779-3784. doi:10.1073/pnas.1019676108

Laurent, Gilles. 2002. "Olfactory Network Dynamics and the Coding of Multidimensional Signals". *Nature Reviews Neuroscience* 3 (11) (November 1): 884-895. doi:10.1038/nrn964

Lee, Seung-Hee, Alex C. Kwan, Siyu Zhang, Victoria Phoumthipphavong, John G. Flannery, Sotiris C. Masmanidis, Hiroki

Taniguchi, *et al*. 2012. "Activation of Specific Interneurons Improves V1 Feature Selectivity and Visual Perception". *Nature* 488 (7411) (August 16): 379-383. doi:10.1038/nature11312

Lisman, J.E., and M.A. Idiart. 1995. "Storage of 7 +/- 2 Short-term Memories in Oscillatory Subcycles". *Science* (New York, N.Y.) 267 (5203) (March 10): 1512-1515.

MacLeod, C.M. 1991. "Half a Century of Research on the Stroop Effect: An Integrative Review". *Psychological Bulletin* 109 (2): 163.

Mathewson, K.E., A. Lleras, D.M. Beck, M. Fabiani, T. Ro, and G. Gratton. 2011. "Pulsed Out of Awareness: EEG Alpha Oscillations Represent a Pulsed-inhibition of Ongoing Cortical Processing". *Frontiers in Psychology* 2.

Mehta, M.R., A.K. Lee, and M.A. Wilson. 2002. "Role of Experience and Oscillations in Transforming a Rate Code into a Temporal Code". *Nature* 417 (6890) (June 13): 741-746. doi:10.1038/ nature00807

Miller, G.A. 1956. "The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information". *Psychological Review* 63 (2) (March): 81-97.

Milner, B. 1963. "Effects of Different Brain Lesions on Card Sorting: The Role of the Frontal Lobes". *Archives of Neurology* 9 (1): 90.

O'Keefe, John, and Michael L. Recce. 1993. "Phase Relationship Between Hippocampal Place Units and the EEG Theta Rhythm". *Hippocampus* 3 (3): 317-330. doi:10.1002/hipo.450030307

Owen, Adrian M., Angela C. Roberts, John R. Hodges, and Trevor W. Robbins. 1993. "Contrasting Mechanisms of Impaired Attentional Set-shifting in Patients with Frontal Lobe Damage or Parkinson's Disease". *Brain* 116 (5) (October 1): 1159-1175. doi:10.1093/brain/116.5.1159

Palva, Satu, and J. Matias Palva. 2011. "Functional Roles of Alpha-band Phase Synchronization in Local and Large-scale Cortical Networks". *Frontiers in Perception Science* 2: 204. doi:10.3389/fpsyg. 2011.00204

Premereur, Elsie, Wim Vanduffel, Pieter R. Roelfsema, and Peter Janssen. 2012. "Frontal Eye Field Microstimulation Induces Task-dependent Gamma Oscillations in the Lateral Intraparietal Area". *Journal of Neurophysiology* 108 (5) (September 1): 1392-1402. doi:10.1152/ jn.00323.2012

Rainer, Gregor, S. Chenchal Rao, and Earl K. Miller. 1999. "Prospective Coding for Objects in Primate Prefrontal Cortex". *The Journal of Neuroscience* 19 (13) (July 1): 5493-5505.

Ray, W.J., and H.W. Cole. 1985. "EEG Alpha Activity Reflects Attentional Demands, and Beta Activity Reflects Emotional and Cognitive Processes". *Science* 228 (4700): 750.

Reynolds, John H., and David J. Heeger. 2009. "The Normalization Model of Attention". *Neuron* 61 (2) (January 29): 168-185. doi:10.1016/j.neuron.2009.01.002

Rigotti, M., D.B.D. Rubin, X.J. Wang, and S. Fusi. 2010. "Internal Representation of Task Rules by Recurrent Dynamics: The Importance of the Diversity of Neural Responses". *Frontiers in Computational Neuroscience* 4.

Rossi, Andrew F., Narcisse P. Bichot, Robert Desimone, and Leslie G. Ungerleider. 2007. "Top-Down Attentional Deficits in Macaques with Lesions of Lateral Prefrontal Cortex". *The Journal of Neuroscience* 27 (42) (October 17): 11306-11314. doi:10.1523/jneurosci.2939-07.2007

Saalmann, Yuri B., Ivan N. Pigarev, and Trichur R. Vidyasagar. 2007. "Neural Mechanisms of Visual Attention: How Top-Down Feedback Highlights Relevant Locations". *Science* 316 (5831) (June 15): 1612-1615. doi:10.1126/science. 1139140

Salinas, Emilio, and Terrence J. Sejnowski. 2000. "Impact of Correlated Synaptic Input on Output Firing Rate and Variability in Simple Neuronal Models". *The Journal of Neuroscience* 20 (16) (August 15): 6193-6209.

Siegel, Markus, Tobias H. Donner, Robert Oostenveld, Pascal Fries, and Andreas K. Engel. 2008. "Neuronal Synchronization Along the Dorsal Visual Pathway Reflects the Focus of Spatial Attention". *Neuron* 60 (4) (November 26): 709-719. doi:10.1016/j.neuron.2008.09.010

Siegel, Markus, and Peter König. 2003. "A Functional Gamma-Band Defined by Stimulus-Dependent Synchronization in Area 18 of Awake Behaving Cats". *The Journal of Neuroscience* 23 (10) (May 15): 4251-4260.

Siegel, Markus, Melissa R. Warden, and Earl K. Miller. 2009. "Phase-dependent Neuronal Coding of Objects in Short-term Memory". *Proceedings of the National Academy of Sciences* 106 (50) (December 15): 21341-21346. doi:10.1073/pnas.0908193106

Tiesinga, P.H.E., J-M. Fellous, J.V. Jos, and T.J. Sejnowski. 2002. "Information Transfer in Entrained Cortical Neurons". *Network: Computation in Neural Systems* 13 (1) (January): 41-66. doi:10.1088/0954-898X/13/1/302

Vijayan, Sujith, and Nancy J. Kopell. 2012. "Thalamic Model of Awake Alpha Oscillations and Implications for Stimulus Processing". *Proceedings of the National Academy of Sciences* 109 (45) (November 6): 18553-18558. doi:10.1073/pnas.1215385109

Vogel, Edward K., and Maro G. Machizawa. 2004. "Neural Activity Predicts Individual Differences in Visual Working Memory Capacity". *Nature* 428 (6984) (April 15): 748-751. doi:10.1038/nature02447

Vogel, Edward K., Andrew W. McCollough, and Maro G. Machizawa. 2005. "Neural Measures Reveal Individual Differences in Controlling Access to Working Memory". *Nature* 438 (7067) (November 24): 500-503. doi:10.1038/nature04171

Wilson, Nathan R., Caroline A. Runyan, Forea L. Wang, and Mriganka Sur. 2012. "Division and Subtraction by Distinct Cortical Inhibitory Networks in Vivo". *Nature* 488 (7411) (August 16): 343-348. doi:10.1038/nature11347

Womelsdorf, Thilo, Jan-Mathijs Schoffelen, Robert Oostenveld, Wolf Singer, Robert Desimone, Andreas K. Engel, and Pascal Fries. 2007. "Modulation of Neuronal Interactions Through Neuronal Synchronization". *Science* 316 (5831) (June 15): 1609-1612. doi:10.1126/science.1139597

The Brain Mechanisms of Conscious Access and Introspection

Stanislas Dehaene

Introduction

γνῶθι σεαυτόν: know thyself. This famous maxim, inscribed in the pronaos of the Apollo temple in Delphi, draws our attention to a remarkable competence of the human brain: the capacity to bring to the forefront of our awareness, not just sensory information from the external world, but also aspects of our inner mental life. Indeed, a characteristic feature of *Homo sapiens sapiens* is that we are conscious of being conscious. A talented painter of introspection, Vladimir Nabokov lyrically summarized, in *Strong Opinions*, the fascinating reflections of this mirror seemingly turned onto itself:

> Being aware of being aware of being… if I not only know that I am but also know that I know it, then I belong to the human species. All the rest follows – the glory of thought, poetry, a vision of the universe. In that respect, the gap between ape and man is immeasurably greater than the one between amoeba and ape.

How does consciousness work? Can it be reduced to the operation of the brain? What are its neurobiological mechanisms? For a long period, these questions were considered beyond the realm of cognitive psychology and neuroscience. Consciousness was considered an unnecessary term. John Broadus Watson forcefully rejected introspection and consciousness from the science of psychology which he sketched in his 1913 manifesto *Psychology as the behaviorist views*:

> Psychology as the behaviorist views it is a purely objective experimental branch of natural science. Its theoretical goal is the prediction and control of behavior. Introspection forms no essential part of its methods, nor is the scientific value of its data dependent upon the readiness with which they lend themselves to interpretation in terms of consciousness.

Although cognitive science rejected behaviorism, the anti-introspection view left a durable mark. During the cognitive revolution (approximately 1960 to 1990), consciousness was barely mentioned, even less studied (with a few major exceptions, e.g. Bisiach, Luzzatti, & Perani, 1979; Frith, 1979; Libet, Alberts, Wright, & Feinstein, 1967; Marcel, 1983; Posner, Snyder, Balota, & Marsh, 1975/2004; Shallice, 1972; Weiskrantz, 1986).

Philosophical approaches failed to shed much light on the problem of how an assembly of nerve cells could produce conscious thoughts. René Descartes, although propounding a materialistic approach to perception, action, emotion and memory, conceived of human consciousness as belonging to an entirely different realm (*res cogitans*) (Descartes, 1648/1937). This dualist position was scientifically unproductive, since it essentially barred any experimental approach. Surprisingly, dualism remained an appealing intuition for some contemporary philosophers (Chalmers, 1996) and even neuroscientists (Eccles, 1994). More recently, other philosophers, capitalizing on their intuitions, introduce additional ill-defined concepts of "qualia", "phenomenal awareness", or "what it is like" to have a certain experience (Block, 1995; Nagel, 1974). Yet others sought a haven in quantum mechanics, in the hope that its mysterious non-deterministic rules would somehow leave room for a conscious observer and free will (Penrose, 1990; Penrose & Hameroff, 1998). It is fair to say, however, that such approaches have not yielded any scientific progress so far, but only theoretical constructs of a highly speculative nature (Eccles, 1994).

It is only in the past twenty years or so that the problem of consciousness recovered its status as a respectable empirical question in experimental psychology and neuroscience. A handful of philosophers (e.g. Churchland, 1986; Dennett, 1991), psychologists (e.g. Baars, 1989; Dehaene & Naccache, 2001), neuropsychologists (e.g. Weiskrantz, 1986) and neuroscientists (e.g. Crick & Koch, 1990; Logothetis, Leopold, & Sheinberg, 1996) argued that consciousness was, first and foremost, a well-defined experimental problem. Indeed, consciousness poses an urgent problem in the clinic where the loss of consciousness in coma, epilepsy or anesthesia is a frequent and yet ill-understood and poorly controlled phenomenon. Fortunately, consciousness can be easily monitored and even manipulated through many different paradigms (e.g. sleep, anesthesia, visual illusions, inattention, confidence reports, etc). The brain mechanisms underlying these manipulations can then be dissected using behavioral measures, neuroimaging and electrophysiology. Animal models of conscious and unconscious behavior may even be conceived (Cowey & Stoerig, 1995).

From this realization emerged a flurry of experimental results and theoretical models. Today, there is both a solid dataset on the brain mechanisms of conscious processing and some convergent theoretical proposals. In this chapter, I will briefly review them (for an in-depth review, see Dehaene & Changeux, 2011).

The multiple meanings of consciousness

Three ingredients permitted a solid line of empirical attack on the problem of consciousness: (1) better definitions of the terms; (2) minimal experimental paradigms; and (3) a careful quantification of introspection. I will consider them in turn.

The word "consciousness", as used in everyday language, is loaded with multiple meanings. Contemporary cognitive neuroscience made progress by recognizing the need to distinguish a minimum of three concepts.

1. **Vigilance**, also called wakefulness, is what varies when we fall asleep or wake up. It relates to the intransitive use of the word "consciousness" in everyday language (as when we say "the patient is still conscious"). It is a necessary but not sufficient condition for conscious access and conscious processing.
2. **Selective attention** is the focusing of mental resources on a subset of the available information. Attention selects some information, separates it from the background, and deepens its processing. Selective attention is typically a non-conscious process that gates access to consciousness.
3. **Conscious access** is the entry of some of the attended information into a second post-perceptual stage of cognitive processing which making it durable, available to many additional cognitive processes, and reportable to others. It relates to the transitive use of the word "consciousness" in everyday language (as when we say "The driver was conscious *of* the red light"). Information which has been consciously accessed can then be submitted to **conscious processing**: it can be channeled, in a typically serial manner, through a series of controlled information-processing stages.

Experiments indicate that the three concepts of vigilance, attention and conscious access are dissociable. For instance, vigilance (or wakefulness) may still exist when conscious access is gone: patients in vegetative state may still have a sleep-wake cycle, but their capacity to access, manipulate and report information is lost. Similarly, attention may exist without conscious access: in the laboratory, we can create conditions in which attention is demonstrably attracted by a flashed picture, and even selectively amplifies it, although the picture remains invisible (Koch & Tsuchiya, 2007; Naccache, Blandin, & Dehaene, 2002). Thus, conscious access is a distinct cognitive entity from both vigilance and selective attention.

Conscious access to sensory information is a simple and well-delimited construct that plays a central role in empirical studies of consciousness (Crick & Koch, 1990; Dehaene, Changeux, Naccache, Sackur, & Sergent,

2006). At any given moment, our brain is bombarded with sensory stimulation, which activates many peripheral sensory areas of the brain. Yet we only gain conscious access to one, or just a few, of these elements of information, while the rest remains unconscious. Conscious access has a limited capacity: if we attend to one object, we may transiently lose consciousness of the surrounding ones. The problem of conscious access consists in understanding what brain mechanisms underlie this limited capacity of consciousness.

There are yet other meanings of consciousness. **Self-consciousness** refers specifically to instances of conscious access in which the information being manipulated or reported is internal to the organism. Multiple aspects of self-consciousness may be distinguished: the capacity to represent our body and its limits; the separation of our actions from those of others (agency); and the formation of a "point of view" on the external world. All of these aspects can be and have been studied experimentally. We understand increasingly well how self-consciousness arises from a combination of brain circuits specializing in the representation of different aspects of our selves (sensory maps of the body, vestibular signals of head stability, programming of intentional movements, etc) (see e.g. Lenggenhager, Tadi, Metzinger, & Blanke, 2007).

There is also **recursive consciousness**, also known as **metacognition**. This is the capacity to "know oneself", i.e. to introspect and obtain information about one's own mental processes. Such information is called "metacognitive" because it provides a higher-order representation of the content, value or quality of some other information represented elsewhere in the system. We rely on metacognition when we evaluate our confidence in a past decision, or when we realize that we do not remember something. A broad array of experimental research, too large to be reviewed here, is available on this topic (Dunlosky & Metcalfe, 2008).

Importantly, research indicates that all of the above aspects of consciousness (vigilance, selective attention, conscious access, self-consciousness and metacognition) are not unique to humans, but are also available to many other animal species such as macaque monkeys. In particular, it is clearly incorrect to think of recursive consciousness as limited to the human species (as Nabokov did in the above citation): there are now well-defined animal models of this ability, in which animals act in ways that indicate some degree of knowledge of their own confidence and fallibility (Terrace & Son, 2009).

Some philosophers consider one last aspect of consciousness as worthy of a separate term: **phenomenal awareness** (Block, 1995; Chalmers, 1996). This term is used to refer to the subjective, feel of conscious experience (also

called *qualia*) – "what it is like" to experience, for instance, a gorgeous sunset or a terrible toothache. Introspectively, there is no doubt that these mental states are real and must be explained. However I share with the philosopher Dan Dennett the view that, as a philosophical concept, phenomenal awareness remains too fuzzily defined to be experimentally useful (Dennett, 2001). Whatever empirical content there is to *qualia* seems to be already covered by the concept of conscious access. A burning sensation, for instance, can be tracked as it makes its way into the brain and becomes transformed from a preconscious sensation in somatosensory cortex to a conscious feeling of pain in the anterior cingulate (Rainville, Duncan, Price, Carrier, & Bushnell, 1997). Whether there is anything left of phenomenal awareness once conscious access is taken care of is highly debatable: the other aspects that philosophers consider as central for the *qualia* concept, such as their ineffable character, remain largely untestable. In the rest of this chapter, I will thus primarily focus on the brain mechanisms of conscious access.

Minimal experimental paradigms for conscious access

The second ingredient that led to the contemporary science of consciousness was the recognition that a broad array of experimental paradigms was available to manipulate conscious access in the lab. With these tools, it became possible to create reproducible states of conscious and unconscious perception (Baars, 1989) (see figure 1, p. 315).

One paradigm is provided by visual illusions such as **binocular rivalry** or **motion-induced blindness.** In these illusions, the stimulus is fixed, and yet the content of consciousness repeatedly changes. In motion-induced blindness, a visible disc, when touched by a cloud of moving dots, transiently vanishes from consciousness at seemingly random moments. In binocular rivalry, two pictures objectively presented to the two eyes alternate in awareness: subjectively, we never see them both at the same time. With such stimuli, it becomes feasible to ask a simple empirical question: Which aspects of brain activity vary in parallel to conscious experience? Neurons in the primary visual cortex typically discharge only in relation to the fixed, objective stimulus, but neurons in higher associative areas of the visual cortex show on and off responses in direct correlation to the subjective reports of visibility and invisibility, making them a neural correlate of conscious access (Logothetis, *et al.*, 1996).

Other visual illusions give scientists complete experimental control over the moment at which sensory stimuli vanish from conscious awareness. One such paradigm is **masking**: a target word or picture is briefly flashed on a computer screen, with an intensity clearly sufficient to make it visible. How-

ever, when the picture is followed, at a short interval, by another such stimulus (the "mask"), it may become totally invisible. Such a stimulus is called **subliminal**, i.e. below the threshold for conscious access. As reviewed below, psychological and brain-imaging experiments indicate that subliminal stimuli continue to be actively processed in the brain at multiple levels: the identity, the meaning, and even the action cued by a subliminal word can be partially activated without awareness. It is now very clear that a great variety of brain regions, located virtually everywhere in the cortex, can operate in a non-conscious mode (with the possible exception of dorsolateral prefrontal cortex). Studies of subliminal processing therefore help delimit what consciousness is *not*.

Inattention offers a third type of experimental paradigm. Here, the subject is temporarily absorbed by a demanding task on a first target T1. During this period, a second target T2 is briefly presented. Under such conditions, the limited capacity of conscious access is such that the second stimulus T2 may fail to be perceived at all, giving rise to **attentional blink** (Raymond, Shapiro, & Arnell, 1992) or **inattentional blindness** (Mack & Rock, 1998). The invisible T2 stimulus is said to be **preconscious**. This term specifically refers to a *temporary* invisibility: a preconscious stimulus, unlike a masked word, may become conscious if it is presented in the absence of any distracting or attention-grabbing thought. Preconscious stimuli are therefore useful in the study of consciousness because the very same stimulus may or may not be conscious at different times, under experimental control.

All of these experimental manipulations provide examples of **minimal contrasts** between conscious and non-conscious processing. In the laboratory, we can create experimental conditions that, in the ideal case, vary *only* in the presence or absence of consciousness. Not only the stimulus itself, but also the participant's responses, can be equated between conscious and non-conscious trials. Indeed, it is possible to exploit the fact that participants often respond at better-than-chance levels to non-conscious stimuli (such non-conscious performance is often called **blindsight**). Contrasting conscious and non-conscious trials in which the same stimulus is presented, and the same correct response is emitted, turns conscious access into a pure experimental variable that can be decorrelated from other input and output contingencies (Lamy, Salti, & Bar-Haim, 2009). The goal of the cognitive neuroscience of consciousness is precisely to understand what types of cognitive processes and brain activity distinguish conscious versus non-conscious trials, or reportable versus non-reportable trials, when everything else is kept identical.

The crucial role of introspection

Not only can sensory stimuli be made to vanish from conscious experience, but it is, in fact, possible to select fixed conditions of stimulation that are just at threshold, such that participants report seeing a stimulus on only half the trials (e.g. Sergent, Baillet, & Dehaene, 2005). By asking participants to report their subjective perception on each trial, we can later sort the trials into "seen" and "unseen", and probe the brain activation differences between them.

This approach illustrates the third key ingredient in the study of consciousness: taking introspection seriously. Introspective reports define the very phenomenon that a science of consciousness purports to study: the subjective, first-person mental states that occupy the mind of a given person and that only he or she knows about. The modern science of consciousness uses numerical scales and other devices to carefully register and quantity subjective introspective reports, such that they can be studied scientifically (Marti, Sackur, Sigman, & Dehaene, 2010; Overgaard, Rote, Mouridsen, & Ramsoy, 2006; Sergent & Dehaene, 2004; Sigman, Sackur, Del Cul, & Dehaene, 2008). The results indicate that illusions can be highly reliable across subjects. This is a crucial fact: although subjectivity is a private and first-person phenomenon, its reports obey psychological laws that are highly reproducible across individuals and can therefore be studied by the standard scientific method (e.g. Marti, *et al.*, 2010).

This realization took some time. As noted in the introduction, introspection has long had a bad reputation in cognitive neuroscience. It was long considered as a poor and unreliable measure that could not be used to found a solid psychological science (Nisbett & Wilson, 1977). This critique, however, conflated two different issues: introspection as a research method, and introspection as raw data. As a research method, introspection cannot be trusted to provide direct information about mental processes. Human subjects often supply inappropriate explanations for their behavior (Johansson, Hall, Sikstrom, & Olsson, 2005). We cannot count on them to tell us how their mind works, precisely because so much of mental computation occurs non-consciously. However, the introspections they provide, however weird or wrong, must still be explained. The correct view is to treat them as raw data in need of an explanation. Visual illusions, in this sense, are "real" phenomena in need of an explanation, and which have the potential to illuminate the mechanisms of consciousness.

Perhaps the best case in point is the "out-of-body" experience in which subjects report a feeling of leaving their body and watching themselves from above. We obviously cannot take them literally – but we can still examine

what brain processes cause this subjective experience. Olaf Blanke's research converges onto a cortical region in the right temporo-parietal junction which, when impaired or electrically perturbed, causes a systematic illusion of self displacement, which can now be systematically reproduced in normal subjects (Blanke, Landis, Spinelli, & Seeck, 2004; Blanke, Ortigue, Landis, & Seeck, 2002) (see Olaf Blanke's chapter in the present volume).

Cognitive signatures of consciousness

With those three ingredients at hand (a focus on conscious access, minimal paradigms contrasting conscious and non-conscious perception, and a careful quantification of introspection), the cognitive psychology and neuroscience of consciousness made huge strides in the past twenty years.

A first axis of research focused on the depth of unconscious processing. Using primarily masked priming and attentional blink paradigms, it was discovered that even stimuli that are totally unconscious can be processed up to a considerable depth (for review, see Kouider & Dehaene, 2007). An unseen picture, word or digit can be identified non-consciously. Even its meaning can be partially extracted. For instance, an unseen emotional word such as "rape", masked below threshold, still activates the amygdala, a brain center involved in fear and other emotions (Naccache, *et al.*, 2005). Even complex operations, such as computing the approximate average of several digits (Van Opstal, de Lange, & Dehaene, 2011) or the combination of multiple decision cues (de Lange, van Gaal, Lamme, & Dehaene, 2011; Dijksterhuis, Bos, Nordgren, & van Baaren, 2006), can unfold without consciousness. The guiding of our movements and the quick inhibition or correction of an inappropriate response also fall within the realm of non-conscious processing (Logan & Crump, 2010; Nieuwenhuis, Ridderinkhof, Blom, Band, & Kok, 2001). The exploration of the limits of non-conscious processing continues to this day, and it is likely that powerful yet non-conscious operations of the brain remain to be discovered. As a rule, we seem to constantly under-estimate the amount of non-conscious processing. It can be said that the vast majority of cognitive operations of the human brain occur without awareness.

While conscious processing thus appears only as the tip of the iceberg, are there cognitive operations can only be deployed when the information is consciously represented? It seems that the answer is positive. With a non-conscious target, cognitive operations can be launched, but they typically do not run to completion. Attaining a firm decision, developing a confident intention, and executing a strategy comprising multiple serial steps, are operations that seem to require conscious perception (de Lange, *et al.*, 2011;

Sackur & Dehaene, 2009). The quality of the extracted information, its durable maintenance and its flexible use in multiple tasks are drastically enhanced on conscious relative to non-conscious trials (Del Cul, Dehaene, Reyes, Bravo, & Slachevsky, 2009).

These data suggest that consciousness is not just an epiphenomenon or an illusion, but fulfills a specific role that may have been positively selected for in evolution: the amplification and global sharing of specific information selected for its likely relevance to the organisms' current goals.

Brain signatures of consciousness

At the neurophysiological level, contrasts between conscious and unconscious stimuli have revealed a number of signatures of consciousness.

Brain imaging techniques have been used, for instance, to track the fate of a flashed visual stimulus such as a word as it enters the retina and, depending on the trial, is or is not consciously perceived. Records of brain activity have revealed that the initial perceptual stages may remain almost strictly identical on conscious and non-conscious trials: the entry of the stimulus into visual areas and its feed-forward propagation into occipital, temporal and parietal cortices can proceed non-consciously (e.g. Sergent, *et al.*, 2005). The brain appears to accumulate evidence about the identity of a subliminal stimulus (Del Cul, Baillet, & Dehaene, 2007), and many specialized areas of the cortex, including motor areas, can receive these unconscious signals and bias their decisions towards the unperceived target (Dehaene, Naccache, *et al.*, 1998; Vorberg, Mattler, Heinecke, Schmidt, & Schwarzbach, 2003).

What seems to be unique to consciousness is a relatively late (~200-300 milliseconds), sudden and non-linear amplification of the incoming activation (Del Cul, *et al.*, 2007). After a brief transition period, the difference between conscious and unconscious trials quickly becomes qualitative, as many areas show a sudden activation ("ignition") only on conscious trials (Dehaene & Changeux, 2005; Del Cul, *et al.*, 2007; Fisch, *et al.*, 2009). When it is conscious, the incoming activation is suddenly amplified and reverberates bidirectionally (bottom-up and top-down) within a large network of distant brain areas, frequently including the original perceptual areas as well higher association cortices in the temporal, parietal and prefrontal lobes. This state of activity is meta-stable and can last for a long duration, long after the original stimulus is gone.

At the surface of the head, conscious ignition is characteristically accompanied by a broad component of the average electro-encephalogram (EEG) called the P300 wave (because its latency is typically 300 milliseconds

or more). The brain generators of the P300 have been shown by intracranial recordings to involve a highly distributed set of nearly-simultaneous active areas including hippocampus and temporal, parietal and frontal association cortices (Gaillard, *et al.*, 2009; Halgren, Marinkovic, & Chauvel, 1998).

Additional signatures of consciousness can be obtained by examining the spontaneous fluctuations of brain signals and whether they index a global, brain-scale state of synchronized activation. A late and distributed burst of local high-frequency activity in the gamma band (>30 Hz), a massive increase in the synchrony between distant brain signals in the beta band (13-30 Hz), and a bidirectional sharing of mutual information and causal links, when occurring in a late time window, all constitute markers of conscious access (Gaillard, *et al.*, 2009).

An important axis of recent research consists in probing the generality of these putative signatures of consciousness (review in Dehaene & Changeux, 2011). Beyond the perception of brief visual stimuli, these markers have begun to be replicated in auditory and tactile perception. Probing these markers during anesthesia and in brain-lesioned patients with loss of consciousness also confirms their tight association with conscious perception.

Importantly, a similar two-stage sequence, with non-conscious focal processing followed by a global synchronous conscious state, has also been observed in studies of conscious access to non-sensory information. For instance, when we are aware of having made an error, a focal and unconscious error-related negativity is followed by a late and global wave, the error positivity, which tightly resembles the sensory P300 (Nieuwenhuis, *et al.*, 2001). A similar sequence can also be evoked by direct brain stimulation: during the conscious state, a magnetic pulse induces activation that propagates to multiple distant brain areas for durations extending beyond 300 ms, while during the anesthetized or sleep state, the same pulse induces only a local activation that quickly dissipates (Ferrarelli, *et al.*, 2010; Massimini, Boly, Casali, Rosanova, & Tononi, 2009). New mathematical measures of information integration or non-linear dimensionality (Velly, *et al.*, 2007) are now being developed to provide improved markers of the global exchange of information across distant areas which characterizes consciousness.

Global workspace theory

My colleagues and I introduced the theory of a Global Neuronal Workspace (GNW) as a putative neurobiological architecture capable of accounting for cognitive and neuroscience observations on unconscious and conscious processing (Dehaene & Naccache, 2001). GNW theory assumes that cortical areas and subcortical nuclei contribute to a great variety of

specialized sub-circuits implementing unconscious and modular "processors" which operate in parallel. Non-conscious stimuli can thus be quickly and efficient processed along automatized or pre-instructed processing routes. However, GNW theory proposes that besides these encapsulated processors, the brain also comprises an architecture which allows a subset of the available information to be globally broadcasted. The GNW breaks the brain's modular organization by allowing selected information to be flexibly routed to various processes of verbal report, evaluation, memory, planning and intentional action (Baars, 1989; Dehaene & Naccache, 2001). Dehaene and Naccache (2001) postulate that "this global availability of information (…) *is* what we subjectively experience as a conscious state".

The hypothetical neurobiological mechanism for global availability is a set of large cortical pyramidal cells with long-range excitatory axons (GNW neurons), together with their relevant thalamo-cortical loops. These cells are present throughout the human cortex, yet they are particularly dense in prefrontal, cingulate, and parietal regions. They form a long-distance network that interconnects associative cortical areas and allows them to flexibly recruit, in a top-down manner, virtually any specialized area. Through their numerous reciprocal connections, GNW neurons are thought to amplify and maintain a specific neural representation for an arbitrary duration, thus keeping it "on line" or "in mind". At any given moment, a conscious content is assumed to be encoded in the sustained activity of a fraction of GNW neurons, the rest being inhibited. The long-distance axons of GNW neurons then broadcast it to many other processors brain-wide. Global broadcasting allows information to be more efficiently processed (because it is no longer confined to a subset of non-conscious circuits, but can be flexibly shared by many cortical processors) and to be verbally reported (because these processors include those involved in formulating verbal messages).

Artificial neuronal networks based on the workspace architecture have been explored in computer simulations (Dehaene & Changeux, 2005; Dehaene, Kerszberg, & Changeux, 1998; Shanahan, 2008; Zylberberg, Fernandez Slezak, Roelfsema, Dehaene, & Sigman, 2010). Their behavior has revealed dynamic electrophysiological phenomena very similar to the above experimental observations. When a brief pulse of sensory stimulation was applied to the model network, activation propagated according to two successive phases: (1) initially, a brief wave of excitation progressed into the simulated hierarchy through feedforward connections, with an amplitude and duration directly related to the initial input; (2) in a second stage, mediated by slower feedback connections, the network entered into a global self-sustained "ignited" state. This ignition was characterized by an increased

power of local cortico-thalamic oscillations in the gamma band, and by an increased synchrony across distant regions. This two-stage dynamics of the computer model reproduced most of the signatures of conscious access that have been empirically observed.

The model easily explains why conscious access exhibits a sharp threshold that separates supra- from sub-liminal stimuli. In GNW theory, the transition to the ignited or "conscious" state can be characterized as a phase transition in network activity. By amplifying its own incoming activity, the GNW exhibits a dynamic threshold with a fast non-linear divergence. Within a few tens of milliseconds, depending on stimulus strength, activity either rises to a high state, or decays to a low state. Even for a fixed stimulus, spontaneous activity and pre-stimulus oscillations impose a stochasticity on global ignition, explaining why the same stimulus can sometimes be perceived and sometimes remain unconscious. Computer simulations also exhibit analogs of the attentional blink and inattentional blindness phenomena: at any given moment, the ignition of the workspace by one cell assembly can prevent the simultaneous conscious access to a second piece of information.

An original feature of the GNW model, absent from many other formal neural network models, is the occurrence of highly structured spontaneous activity (Dehaene & Changeux, 2005). Just like real neurons, the simulated GNW neurons can fire spontaneously, with a fringe of variability, even in the absence of external inputs. In a GNW architecture, this spontaneous activity propagates in a top-down manner, starting from the highest hierarchical levels of the simulation, to form globally synchronized ignited states. The dynamics of such networks is thus characterized by a constant flow or "stream" of individual coherent episodes of variable duration In more complex network architectures, this stochastic activity can be shaped by reward signals in order to achieve a defined goal state, such as solving a logical problem (Dehaene & Changeux, 2000). These simulations provide a preliminary account of how higher cortical areas spontaneously activate in a coordinated manner during conscious effortful tasks.

In summary, the theoretical proposal is that conscious access corresponds to the selection and temporary maintenance of information encoded in the sustained activity of a distributed network of neurons with long-distance axons (the Global Neuronal Workspace). The GNW theory accounts for at least three aspects of subjective experience: (1) individuality: the same stimulus may or may not lead to conscious ignition, and whether such ignition occurs, in a given brain, is a stochastic event unique to each individual; (2) durability: thanks to its reverberating self-connectivity, the GNW network can maintain information "in mind" for an arbitrary duration, long after

the actual sensory stimulation has vanished; (3) autonomy: the shaping of spontaneous activity by GNW circuits leads to the stochastic endogenous generation of a series of activation patterns, potentially accounted for the never-ending "stream of consciousness".

Towards clinical applications

The discovery of the brain mechanisms of consciousness is not just an intellectual exercise. Our research is strongly motivated by the need to provide better experimental and conceptual tools to clinicians. Every year, due to stroke, head trauma or hypoxia, thousands of patients lose consciousness. The current clinical classification distinguishes several states:

- Brain death: complete and irreversible absence of brain function, marked by the durable absence of any detectable electro-encephalogram (EEG) and brain stem reflexes, which cannot be explained by hypothermia or drugs.
- Coma: prolonged loss of the capacity to be aroused, typically accompanied by slow-wave EEG and a variable preservation of cranial nerve and brain stem reflexes.
- Vegetative state: preserved sleep-wake cycle, yet with a total lack of responsiveness and voluntary action.
- Minimally conscious state: presence of rare, inconsistent, and limited signs of understanding and voluntary responding.
- Locked-in syndrome: fully preserved awakening and awareness, yet with complete or near-complete incapacity to report it due to paralysis (eye motion can be preserved).

Clinical scales, unfortunately, are not devoid of ambiguity. Brain imaging indicates that a few patients in apparent clinical vegetative state may, in fact, present residual consciousness. They exhibit complex and essentially normal cortical responses to speech, as well as a capacity to follow instructions such as "imagine visiting your apartment" (Owen, *et al.*, 2006). Functional magnetic resonance imaging (fMRI) can even be used to communicate with such patients, using very indirect instructions such as "if you want to respond yes, imagine visiting your apartment, otherwise imagine playing tennis", and monitoring the activity of the corresponding brain networks as a proxy for the patient's response (Monti, *et al.*, 2010).

In the near future, there is great hope that the current progress in understanding the signatures of consciousness will lead to easier and more theoretically justified clinical tools. Compared to fMRI, EEG should provide a simpler means to detect rare cases of residual awareness, but also to

improve the diagnosis of all coma and vegetative state patients and to sharpen the prediction of their awakening and future cognitive state. EEG is already used to monitor the depth of propagation of auditory signals in order to predict the recovery of coma patients (Fischer, Luaute, Adeleine, & Morlet, 2004; Kane, Curry, Butler, & Cummins, 1993). In our laboratory, we have developed a paradigm to specifically isolate the P300 wave which is evoked in response to novel auditory signals (Bekinschtein, *et al.*, 2009). In agreement with research in normal subjects, the detection of this wave facilitates the diagnosis of patients with residual awareness and/or imminent recovery (Faugeras, *et al.*, 2011). It is also possible to stimulate the brain with a pulse of external, magnetically induced activity. Again, as theoretically predicted, the duration, complexity, and distance of the propagation to other cortical sites indexes the recovery of consciousness (Rosanova, *et al.*, 2012). Other signatures, such as mathematical indices of the long-distance synchrony between brain areas, may prove to be even more sensitive (King, Dehaene, Naccache *et al.*, in preparation).

Conclusion

The subjective aspects of conscious experience no longer lie beyond the realm of an objective scientific inquiry. On the contrary, a solid body of scientific evidence links consciousness to specific cognitive computations and to the physical state of networks of neurons. Advances in brain imaging now make it possible to reliably detect electrophysiological signatures of consciousness. These signatures can be used to decide, with above-chance accuracy, whether a normal person is or is not aware of a given stimulus, or whether a patient still presents a residual form of consciousness.

While these advances are significant, it should be stressed that they concern primarily the simplest sense of the term "consciousness": the ability to gain conscious access to some information. The brain mechanisms underlying the capacity for self-consciousness (knowing that we know) are only starting to be studied with similar methods (e.g. Fleming, Weil, Nagy, Dolan, & Rees, 2010).

References

Baars, B.J. (1989). *A cognitive theory of consciousness*. Cambridge, Mass.: Cambridge University Press.

Bekinschtein, T.A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., & Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. *Proc Natl Acad Sci U S A*, 106(5), 1672-1677.

Bisiach, E., Luzzatti, C., & Perani, D. (1979). Unilateral neglect representational schema

and consciousness. *Brain*, 102, 609-618.
Blanke, O., Landis, T., Spinelli, L., & Seeck, M. (2004). Out-of-body experience and autoscopy of neurological origin. *Brain*, 127(Pt 2), 243-258.
Blanke, O., Ortigue, S., Landis, T., & Seeck, M. (2002). Stimulating illusory own-body perceptions. *Nature*, 419(6904), 269-270.
Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-287.
Chalmers, D. (1996). *The conscious mind*. New York: Oxford University Press.
Churchland, P.S. (1986). *Neurophilosophy: toward a unified understanding of the mind/brain*. Cambridge MA: MIT Press.
Cowey, A., & Stoerig, P. (1995). Blindsight in monkeys. *Nature*, 373(6511), 247-249.
Crick, F., & Koch, C. (1990). Toward a neurobiological theory of consciousness. *Seminars in Neuroscience*, 2, 263-275.
de Lange, F.P., van Gaal, S., Lamme, V.A., & Dehaene, S. (2011). How awareness changes the relative weights of evidence during human decision-making. *PLoS Biol*, 9(11), e1001203.
Dehaene, S., & Changeux, J.P. (2000). Reward-dependent learning in neuronal networks for planning and decision making. *Prog Brain Res*, 126, 217-229.
Dehaene, S., & Changeux, J.P. (2005). Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. *PLoS Biol*, 3(5), e141.
Dehaene, S., & Changeux, J.P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70, 200-227.
Dehaene, S., Changeux, J.P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn Sci*, 10(5), 204-211.
Dehaene, S., Kerszberg, M., & Changeux, J.P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc Natl Acad Sci U S A*, 95(24), 14529-14534.
Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79, 1-37.
Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., *et al.* (1998). Imaging unconscious semantic priming. *Nature*, 395, 597-600.
Del Cul, A., Baillet, S., & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biol*, 5(10), e260.
Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, 132, 2531–2540.
Dennett, D. (1991). *Consciousness explained*. London: Penguin.
Dennett, D. (2001). Are we explaining consciousness yet? *Cognition*, 79(1-2), 221-237.
Descartes, R. (1648/1937). *Traité de l'homme Descartes: Oeuvres et lettres*. Paris: Gallimard.
Dijksterhuis, A., Bos, M.W., Nordgren, L.F., & van Baaren, R.B. (2006). On making the right choice: the deliberation-without-attention effect. *Science*, 311(5763), 1005-1007.
Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*: Sage Publications, Inc.
Eccles, J.C. (1994). *How the self controls its brain*. New York: Springer Verlag.
Faugeras, F., Rohaut, B., Weiss, N., Bekinschtein, T. A., Galanaud, D., Puybasset, L., *et al.* (2011). Probing consciousness with event-related potentials in the vegetative state. *Neurology*, 77(3), 264-268.
Ferrarelli, F., Massimini, M., Sarasso, S., Casali, A., Riedner, B.A., Angelini, G., *et al.* (2010). Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness. *Proc Natl Acad Sci U S A*, 107(6), 2681-2686.

Fisch, L., Privman, E., Ramot, M., Harel, M., Nir, Y., Kipervasser, S., *et al.* (2009). Neural "Ignition": Enhanced Activation Linked to Perceptual Awareness in Human Ventral Stream Visual Cortex. *Neuron*, 64, 562-574.

Fischer, C., Luaute, J., Adeleine, P., & Morlet, D. (2004). Predictive value of sensory and cognitive evoked potentials for awakening from coma. *Neurology*, 63(4), 669-673.

Fleming, S.M., Weil, R.S., Nagy, Z., Dolan, R.J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, 329(5998), 1541-1543.

Frith, C.D. (1979). Consciousness, information processing and schizophrenia. *Br J Psychiatry*, 134, 225-235.

Gaillard, R., Dehaene, S., Adam, C., Clemenceau, S., Hasboun, D., Baulac, M., *et al.* (2009). Converging intracranial markers of conscious access. *PLoS Biol*, 7(3), e61.

Halgren, E., Marinkovic, K., & Chauvel, P. (1998). Generators of the late cognitive potentials in auditory and visual oddball tasks. *Electroencephalogr Clin Neurophysiol*, 106(2), 156-164.

Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310(5745), 116-119.

Kane, N.M., Curry, S.H., Butler, S.R., & Cummins, B.H. (1993). Electrophysiological indicator of awakening from coma. *Lancet*, 341(8846), 688.

Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends Cogn Sci*, 11(1), 16-22.

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philos Trans R Soc Lond B Biol Sci*, 362(1481), 857-875.

Lamy, D., Salti, M., & Bar-Haim, Y. (2009). Neural Correlates of Subjective Awareness and Unconscious Processing: An ERP Study. *J Cogn Neurosci*, 21(7), 1435-1446.

Lenggenhager, B., Tadi, T., Metzinger, T., & Blanke, O. (2007). Video ergo sum: manipulating bodily self-consciousness. *Science*, 317(5841), 1096-1099.

Libet, B., Alberts, W.W., Wright, E.W., Jr., & Feinstein, B. (1967). Responses of human somatosensory cortex to stimuli below threshold for conscious sensation. *Science*, 158(808), 1597-1600.

Logan, G.D., & Crump, M.J. (2010). Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science*, 330(6004), 683-686.

Logothetis, N.K., Leopold, D.A., & Sheinberg, D.L. (1996). What is rivalling during binocular rivalry? *Nature*, 380(6575), 621-624.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, Mass.: MIT Press.

Marcel, A.J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, 15, 197-237.

Marti, S., Sackur, J., Sigman, M., & Dehaene, S. (2010). Mapping introspection's blind spot: Reconstruction of dual-task phenomenology using quantified introspection. *Cognition*, 115(2), 303-313.

Massimini, M., Boly, M., Casali, A., Rosanova, M., & Tononi, G. (2009). A perturbational approach for evaluating the brain's capacity for consciousness. *Prog Brain Res*, 177, 201-214.

Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., Boly, M., Pickard, J. D., Tshibanda, L., *et al.* (2010). Willful Modulation of Brain Activity in Disorders of Consciousness. *N Engl J Med*, 362(7), 579-589.

Naccache, L., Blandin, E., & Dehaene, S. (2002). Unconscious masked priming depends on temporal attention. *Psychological Science*, 13, 416-424.

Naccache, L., Gaillard, R., Adam, C., Has-

boun, D., Clémenceau, S., Baulac, M., *et al*. (2005). A direct intracranial record of emotions evoked by subliminal words. *Proc Natl Acad Sci U S A*, 102, 7713-7717.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review* 83(4), 435-450.

Nieuwenhuis, S., Ridderinkhof, K.R., Blom, J., Band, G.P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology*, 38(5), 752-760.

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, 84(3), 231-259.

Overgaard, M., Rote, J., Mouridsen, K., & Ramsoy, T.Z. (2006). Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task. *Conscious Cogn*.

Owen, A.M., Coleman, M.R., Boly, M., Davis, M.H., Laureys, S., & Pickard, J.D. (2006). Detecting awareness in the vegetative state. *Science*, 313(5792), 1402.

Penrose, R. (1990). *The emperor's new mind. Concerning Computers, Minds, and the Laws of Physics*. London: Vintage books.

Penrose, R., & Hameroff, S. (1998). The Penrose-Hameroff "Orch OR" model of consciousness. *Philosophical Transactions Royal Society London* (A), 356, 1869-1896.

Posner, M.I., Snyder, C.R.R., Balota, D.A., & Marsh, E.J. (1975/2004). *Attention and Cognitive Control Cognitive psychology: Key readings*. (pp. 205-223). New York, NY US: Psychology Press.

Rainville, P., Duncan, G.H., Price, D.D., Carrier, B., & Bushnell, M.C. (1997). Pain affect encoded in human anterior cingulate but not somatosensory cortex. *Science*, 277(5328), 968-971.

Raymond, J.E., Shapiro, K.L., & Arnell, K.M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? *J Exp Psychol Hum Percept Perform*, 18(3), 849-860.

Rosanova, M., Gosseries, O., Casarotto, S., Boly, M., Casali, A.G., Bruno, M.A., *et al*. (2012). Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. *Brain*, 135(Pt 4), 1308-1320.

Sackur, J., & Dehaene, S. (2009). The cognitive architecture for chaining of two mental operations. *Cognition*, 111(2), 187-211.

Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nat Neurosci*, 8(10), 1391-1400.

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychol Sci*, 15(11), 720-728.

Shallice, T. (1972). Dual functions of consciousness. *Psychol Rev*, 79(5), 383-393.

Shanahan, M. (2008). A spiking neuron model of cortical broadcast and competition. *Conscious Cogn*, 17(1), 288-303.

Sigman, M., Sackur, J., Del Cul, A., & Dehaene, S. (2008). Illusory displacement due to object substitution near the consciousness threshold. *J Vis*, 8(1), 13 11-10.

Terrace, H.S., & Son, L.K. (2009). Comparative metacognition. Curr Opin Neurobiol, 19(1), 67-74.

Van Opstal, F., de Lange, F.P., & Dehaene, S. (2011). Rapid parallel semantic processing of numbers without awareness. *Cognition*.

Velly, L.J., Rey, M.F., Bruder, N.J., Gouvitsos, F.A., Witjas, T., Regis, J.M., *et al*. (2007). Differential dynamic of action on cortical and subcortical structures of anesthetic agents during induction of anesthesia. *Anesthesiology*, 107(2), 202-212.

Vorberg, D., Mattler, U., Heinecke, A., Schmidt, T., & Schwarzbach, J. (2003). Different time courses for visual perception and action priming. *Proc Natl Acad Sci U S A*, 100(10), 6275-6280.

Weiskrantz, L. (1986). *Blindsight: A Case Study and Its Implications*. Oxford: Clarendon Press.

Zylberberg, A., Fernandez Slezak, D., Roelfsema, P. R., Dehaene, S., & Sigman, M. (2010). The brain's router: a cortical network model of serial processing in the primate brain. *PLoS Comput Biol*, 6(4), e1000765.

Consciousness and Self-Consciousness in Favour of a Pragmatic Dualism in the Philosophy of Mind

Jürgen Mittelstrass

Consciousness and self-consciousness, or self-understanding, are among the central concepts of philosophy in its European tradition – like nature and reason. Man is the animal which is conscious of its doings, its cognition and its situation in the world and which is able to relate, at the same time, to this consciousness cognitively and reflectively. Philosophy addresses these relationships in the domain of epistemology, but increasingly so, too, does natural science in the form of (cognitive) neuroscience and, in particular, brain research. The natural sciences are getting involved with philosophical conceptions, but philosophy is getting equally involved with scientific procedures and results. This latter proceeds by way of the philosophy of science (of the neurosciences), as well as by way of more anthropological approaches. Knowledge about man is scientific and philosophical (epistemological and anthropological) at the same time. This gives rise to conflicts, especially when scientific knowledge claims to encompass all knowledge of man. Everything that is the case is amenable to scientific explanation – thus the fundamental conviction of the natural sciences. Is this also the case with consciousness and self-consciousness?

As far as the natural sciences are concerned, the objective is to explain how consciousness works from the physiological point of view and what capacities it has – in the words of the brain researcher: "to attribute a large part of our cognitive and motoric capacities to the brain and to conceive of deficiencies of these functions as entirely standard organic diseases".[1] As far as philosophy is concerned, the objective is to explain from the epistemological point of view how consciousness is mirrored in its cognition and its other subjective performances. The cognition and the reflection of this mirroring, in turn, is self-consciousness. The natural sciences and philosophy are also at loggerheads about this topic, self-consciousness. Is it possible to 'explain' self-consciousness just like it is possible to explain consciousness,

[1] W. Singer, "Einführung", in: *Gehirn und Bewusstsein*, Heidelberg and Berlin and Oxford 1994, p. VII, my translation.

or is it of a different kind? The natural sciences say no, philosophy says yes, and tries to express this in the conceptions used – for instance the concept of the *self* or the *ego*. In what follows, divided in three parts, I present a critical analysis of the philosophical and scientific approaches, respectively, but I shall begin with a short recollection of the career of the concepts of consciousness and self-consciousness.

1. The ego

Consciousness has always been understood as a cognition that is not just displayed in mere behaviour but which articulates itself as a 'consciousness of something'. To articulate here means: to differentiate, to conceptualize, to assess, to connect the perceived (what is given in perception) with the constructed (what is determined in thought). In the process consciousness becomes, in modern terminology, a property of the mental or of mental states and conditions. This finds its epistemological expression in the Aristotelian concept of thinking, *noesis*, complementing a mere perception with an aspect of intentionality: consciousness as an action directed towards a certain matter of fact or as object-related knowing, which articulates itself linguistically (conceptually).

In contrast to the concept of consciousness as object-related perception, the concept of self-consciousness means the perception of an object-related perception (and other subjective performances), a consciousness, thus, which becomes self-reflective and to this extent also may be understood as condition of all cognition in its philosophic and scientific forms. Also this aspect may already be found in Aristotle, namely in the phrase "thinking of thinking" (νόησις νοήσεως),[2] where Aristotle assigns the concept of a pure self-consciousness to the concept of a pure reason, which turns out to be a condition of philosophy and science.[3] In Descartes, this issue becomes the fundamental principle of his philosophy of science and metaphysics; it also does this in the further development of both rationalist and empiricist epistemological perspectives. In Leibniz, for instance, the perceptions of the (rational) monads (souls) are apperceptions, defined as reflective consciousness, in Locke, the 'ideas of reflection' are the result of perceiving one's own cognitions. At the same time, the concept of self-consciousness is related to the concept of an I-substance, the *ego*, which in Kant – where all currents of

[2] *Met.* 9.1074b34.

[3] On this and the further history of the concept of self-consciousness, see C.F. Gethmann, "Selbstbewusstsein", in: J. Mittelstrass (Ed.), *Enzyklopaedie Philosophie und Wissenschaftstheorie*, vol. III, Stuttgart and Weimar 1995, pp. 755-759.

philosophical tradition meet and are put on new, critical foundations – in turn finds its transcendental reformulation.[4]

In Kant's terminology, the *ego*, in reference to itself, perceives itself as *appearance*, not as substance, and it is empirically given only in this sense. With the concept of the *transcendental subject*, this idea gains epistemological significance as the principle of unity of knowledge and things: "For inner experience in general and its possibility, or perception in general and its relation to another perception, without any particular distinction or empirical determination being given in it, cannot be regarded as empirical cognition, but must be regarded as cognition of the empirical in general, and belongs to the investigation of the possibility of every experience, which is of course transcendental".[5] Self-consciousness, in this sense, again means the ability of the subject to refer, with the intention of knowing, to its own object-related knowing. Kant uses here the well-known formula: "The *I think* must *be able* to accompany all my representations", followed by the explanation: "for otherwise something would be represented in me that could not be thought at all, which is as much as to say that the representation would either be impossible or else at least would be nothing for me".[6] From here the further development leads on the one hand, against the idealistic theory of the *ego* and self-consciousness, to a philosophy of concrete subjectivity, to a phenomenology of *ego*-perceptions and, on the other hand, to psychological theories as well as analytical approaches.

In this development, the concept of the *self* is either identical with the concept of the *ego* or, in contrast to this concept, emphasizes the more phenomenological aspects of individual forms of existence and self-understanding. For Leibniz, it was self-reflection that makes it possible to say '*I*', Kant distinguishes between a *determining self* (thought) and a *determinable self* (the thinking subject) without associating this distinction with any distinctions between *ego* and *self*. In contrast to the identity of the *ego* by which more abstract aspects are emphasized – these aspects still influenced Husserl's concept

[4] The following is closely based on an earlier account: J. Mittelstrass, "Le soi philosophique et l'identité de la rationalité philosophique", in: E.D. Carosella *et al.* (Eds.), *L'identité changeante de l'individu. La constante construction du Soi*, Paris 2008, pp. 203-212 (especially pp. 207-210); English version: "The Philosophical Self and the Identity of Philosophical Rationality", in: J. Chr. Heilinger *et al.* (Eds.), *Individualitaet und Selbstbestimmung*, Berlin 2009, pp. 55-61 (especially pp. 58-59).

[5] *Critique of Pure Reason* B 401 (translation from *Critique of Pure Reason* [transl. and ed. P. Guyer and A.W. Wood], Cambridge 1998, p. 412).

[6] *Critique of Pure Reason* B 132-133 (translation from *Critique of Pure Reason* [see footnote 5], p. 246).

of the transcendental *ego* –, the concept of the *self*, for instance in Heidegger, aims at the phenomenal variety of personal identity (the 'authentically existing self',[7] '*Dasein'* [existence] as 'being-within-the-world').

As already pointed out, the concept of reflection is closely related to the concepts of the *ego* and the *self* or rather the concept of self-consciousness. This concept stands for the self-ascertainment of the *ego* or the *self*, in epistemological terms for the '*I think'* which, according to Kant, accompanies all judgements and all activities of the understanding. Thought, in this respect, is self-reflexive by nature and, correspondingly, so is the cogitating *ego* and the cogitating *self*.

A further step is made by Fichte when he says that in thinking the *ego* creates itself. Here, the *ego* is perceived as absolute *ego*, as pure self-performance that even constitutes in itself the difference between *ego* and non-*ego*, i.e. nature. With this, a logical level is reached where it is no longer the constitution of the individual that is at stake (according to the related concepts of *ego*-identity and *self*-identity) but, as in Kant in an epistemological framework, the constitution of a philosophical *ego* or philosophical *self* – in Kant's terminology: the constitution of a transcendental subject. The identity of this subject consists in the fact that it is neither the particular (empirical) subject nor the universal (theoretical) subject, but the condition of both. In this sense, Wittgenstein writes: "The subject does not belong to the world, but it is a limit of the world".[8] Wittgenstein here refers to the individual subject, but his statement is also precisely true in view of the fact that the acting *ego* (Kant: the determining *ego*), in its performances, cannot be grasped theoretically.[9] Just this is expressed in the concepts of reflectivity and the transcendental.

What is expressed by the terminology of *ego* or *self*, as well as in the expression that the acting or determining *ego* cannot be grasped theoretically, marks the frontier at which the natural sciences in the figure of brain research and philosophy in the form of epistemology stand opposed to one another in critical conflict.

2. Science and the philosophy of mind

Where a science claims to explain everything or at least a great variety of many different things with the same method, it either becomes dogmatic

[7] *Sein und Zeit*, 8th ed., Tuebingen 1957, p. 130.

[8] *Tractatus logico-philosophicus* 5.632 (translation from: *Tractatus logico-philosophicus. With an Introduction by Bertrand Russell*, London 1922, 1947, p.151).

[9] See K. Lorenz, "Identitaet", in: J. Mittelstrass (Ed.), *Enzyklopaedie Philosophie und Wissenschaftstheorie*, 2nd ed. vol. III, Stuttgart and Weimar 2008, pp. 530-534.

or lets itself be guided by a methodological and theoretical paradigm that makes a claim to universality. The kind of physicalism propounded in the context of Logical Empiricism is an example. It says that all knowing may be expressed using the language of physics and, what is more, that all scientific theories are ultimately reducible to theories of physics. This is an expression of the covert or open *reductionism* of the natural sciences, that is, the programmatic idea of tracing back scientific explanations to uniform notions of a conceptual, methodological and theoretical kind, aiming at a universal explanatory competence. Philosophically, in the traditional sense, this is a variety of *monism,* which in this case leads to a *naturalism*. Naturalism claims that scientific claims of validity are to be traced back to natural facts (that is, facts captured by science), which implies a naturalizing of cognition itself. In the neurosciences, especially in brain research, such a claim is based on the thesis, that also characterizes physicalism, of the causal closure of the physical world. That would include, accordingly, also the spheres of consciousness and self-consciousness.

In modern philosophy, more specifically in the philosophy of mind, this corresponds to so-called Eliminative Materialism and the so-called Identity Theory, especially in the theoretical variety of type-identity. According to Eliminative Materialism, cognitive psychology and folk psychology will be replaced, materially and conceptually, by progress in neurophysiology;[10] according to the Identity Theory, mental states and processes are identical to states and processes of the human brain.[11] Following the theory of type-identity, this includes the claim of a (future) reducibility of psychological statements to neurophysiological laws.[12] Thus Eliminative Materialism and the Identity Theory represent the philosophical foundations of the reductionist claims of (parts of) brain research, to be (or to become) the 'whole' explanation in matters of consciousness and self-consciousness.

[10] See P.M. Churchland, "Eliminative Materialism and the Propositional Attitudes", *The Journal of Philosophy* 78 (1981), pp. 67-90; P.S. Churchland, *Neurophilosophy. Toward a Unified Science of the Mind/Brain*, Cambridge Mass. and London 1986, 1988.

[11] See H. Feigl, *The 'Mental' and the 'Physical'. The Essay [1958] and a Postscript*, Minneapolis Minn. 1967; J.J.C. Smart, "The Mind/Brain Identity Theory", in: *The Stanford Encyclopedia of Philosophy (Fall 2011 Edition)*, E.N. Zalta (Ed.), URL http://plato.stanford.edu/archives/fall2011/entries/mind-identity/

[12] For an account and discussion of various theories in the domain of the philosophy of mind see M. Carrier and J. Mittelstrass, *Mind, Brain, Behavior. The Mind-Body Problem and the Philosophy of Psychology*, Berlin and New York 1991, also M. Carrier, "Philosophy of mind", in: J. Mittelstrass (Ed.), *Enzyklopädie Philosophie und Wissenschaftstheorie*, vol. III, pp. 220-226.

By no means, however, have all issues been resolved, as far as science or philosophy is concerned. In fact, philosophy of mind leads to a trilemma, which according to B. Falkenburg may be represented by the following three theses: "*Radical diversity:* mental phenomena, that is, the mental states, processes or events which we experience, are not physical. In other words, they are *strictly different* from all physical phenomena. *Mental causation:* Mental phenomena may *cause* physical phenomena, that is, our conscious intentions may cause bodily movements in the external world. *Causal closure:* The domain of physical phenomena is causally closed, that is, physical states, processes and events have only *physical* but no non-physical causes".[13]

The third thesis is the thesis adopted by large parts of the cognitive neurosciences. It corresponds, in the philosophy of mind, to the Identity Theory understood as a theory of type-identity and thus philosophically represents a reductionist and naturalist worldview. But it is built on sand. In fact, the cognitive neurosciences, and in particular the brain sciences, have not managed to demonstrate causal (neuronal) mechanisms which could explain consciousness, let alone self-consciousness – and thus the interactions of brain and mind. "Consciousness is and remains mysterious".[14] By making reference to the thesis of causal closure of nature, this position proves to be unfounded from the scientific point of view, and so does the thesis of determinism it endorses – there is no uniform principle of causation in modern physics which could serve as a foundation for a strict determinism – and from the philosophical point of view it proves to be metaphysical.[15] For instance in quantum theory, probabilistic state descriptions of micro-objects lead to indeterminist theories; in philosophy, metaphysical points of view give way to conceptions from the philosophy of science and language. But thus the trilemma mentioned above is losing its philosophical significance: the first and the second thesis remain philosophically viable, the third does not. It is the language of the cognitive neurosciences in particular that give the false impression of neuronal determinism and thus a worldview which seemingly does not leave space for the distinction between physical and non-physical phenomena anymore.[16]

[13] B. Falkenburg, *Mythos Determinismus. Wieviel erklaert uns die Hirnforschung?,* Berlin and Heidelberg 2012, pp. 28-29, my translation.

[14] B. Falkenburg, op. cit., p. 379.

[15] See B. Falkenburg, op. cit., pp. 370-378.

[16] On this and for a critique of this worldview, see M.R. Bennett and P.M. St. Hacker, *Philosophical Foundations of Neuroscience*, Malden Mass. and Oxford, and Carlton 2003; also P. Janich, *Kein neues Menschenbild. Zur Sprache der Hirnforschung,* Frankfurt 2009. Janich draws attention to the consequences of an alleged neuronal determinism, namely

Mediating between the positions of strict neuronal determinism and that of metaphysical dualism as supported by the rationalist tradition, but also, for instance, by Popper and Eccles[17] (the independent existence of mental and physical states), today there also is the conception of *pragmatic dualism*.[18] This view leaves the theoretical possibility of body-mind-identity open, but shows how psychophysical interactionism is the most convincing option, given the current state of science, also avoiding the above-mentioned trilemma. 'Consciousness', 'self-consciousness', and '*ego*' are dualistic terms; they cannot be formed in a monistic conception. They are titles of a specifically philosophical way of orientating oneself in thought and through thought without blocking the way to science. That means: the conception of pragmatic dualism neither anticipates future scientific developments nor does it exclude any particular scientific developments, nor does it simply adopt uncritically earlier metaphysical positions developed in the context of the so-called mind-body problem.

3. Free will

Dualist and monist views of consciousness and self-consciousness clash nowhere as vehemently as they do on the question of *free will*. Not just for philosophical reason, but also for common sense, conscious decisions are the causes of actions. First we decide, then we act; first there is consciousness, then there is the action. In opposition to that stands the thesis of the neurophysiologist that consciousness is a merely *interpreting* and not an acting authority: it is not consciousness that moves; other, physical and mental circumstances move. On that view the actual causes of actions are connected to physical and psychological mechanisms, which are not amenable to introspection, in which consciousness is looking at itself, as it were, and thus not amenable to conscious experience. But this would mean that consciousness would not have a privileged access to the originating conditions of an action; instead it would find itself in the position of an external spectator, as it were. It is not influencing decisions, but only registering them and

the obligation to attribute sense and reference to neuronal states and processes themselves: "When a brain researcher makes the claim that 'ultimately' the meaning and validity of linguistic communication should be explicable via neuronal functions, which already have meaning and validity, he is fudging. He is cheating his way to acceptance of his claim by already attributing the properties of meaningful speech to the material building blocks of his models of the brain" (*op. cit.*, p. 73, my translation).

[17] K.R. Popper and J.C. Eccles, *The Self and Its Brain*, Berlin and London and New York 1977, 1985.

[18] M. Carrier and J. Mittelstrass, *Mind, Brain, Behavior* (footnote 12).

dressing them in a meaning that appears plausible. It invents good reasons, which however do not have anything to do with the actual causes. It is quite clear that this conception appears rather absurd, considered against the background of our self-experience and our self-understanding. How could we be able to imagine that consciousness, experienced subjectively as the source of our activities should be causally ineffective.

Already for Greek thought, the concept of the will marks the passage from prudence or deliberation to action; the freedom of the will accordingly meant the space of action between doing and not doing something. The issue was the concept of well-founded willing, not the search after some mysterious substance in body, soul or reason. A person whose actions are guided by rational considerations is free, or rightly called the possessor of a free will. The further philosophical development went a different way. The will is now considered a separate source of action, next to reason. By leaving behind the prudence-model of the will problems of determinism arise for the first time. They do not concern the idea of free action but rather the idea of free willing, that is, the idea of the free will as an uncaused willing. The thesis is: We may do or not do what we will; but we cannot will or not will whatever we will. This is how Schopenhauer's writings and his thesis of the world as will and representation is to be read.

The right keyword gets voiced (as so often is the case) with Kant. Next to a 'causality according to the laws of nature' there is a 'causality of freedom'. The issue is again (just like in Greek thought) *freedom of action*, not some sort of substance, called freedom or free will, and the problem of well-founded (rational) action, in traditional terminology: the problem of a rational (or good) will. 'Causality of freedom' – this is, in other words, the capacity to act according to *principles*. The point is demands (in the sense of principles) addressed to ourselves, and the realization of these demands. Everybody, including the natural scientist, understands what is meant by this, even if it is a 'causality according to the laws of nature' and not a 'causality of freedom' that he is looking for as scientist. Using the terminology of freedom and the will: we are free in formulating the principles and in (willingly) following them. It is not the free will that is the problem, but the rational will (and thus the determination of the will as self-determination), articulated in the demand to act according to rational reasons.[19]

[19] See J. Mittelstrass, "Der arme Wille. Zur Leidensgeschichte des Willens in der Philosophie", in: H. Heckhausen *et al.* (Eds.), *Jenseits des Rubikon. Der Wille in den Humanwissenschaften*, Berlin etc. 1987, pp. 33–48, also in: J. Mittelstrass, *Der Flug der Eule. Von der Vernunft der Wissenschaft und der Aufgabe der Philosophie*, Frankfurt 1989, pp. 142–163.

When people see this differently and take the causal closure of nature for granted, as natural scientists do, it is primarily *semantic* problems that cloud the view on the differences. In this particular case, the scientific side is unable to imagine anything else than that the non-scientific positions supported by others will eventually join them in believing that the decision between '(free will) exists' and '(free will) does not exist' will be made on experimental grounds, so to speak, that semantic problems are mere pseudo-problems. But the untenable position of causal closure and the arguments brought forward for pragmatic dualism render it obvious that they are not pseudo-problems. It is, in any case, also true that if the claim that the will is causally determined throughout were true, then that claim itself and its claim to validity would be determined causally, that is, via natural causalities. Or to put it differently: a world without freedom would be a world without *reasons*, and for this reason – this is often overlooked by the reductionist and naturalist approaches – a world without science. Hence: science itself is the most beautiful refutation of the scientific negation of a free will.[20]

So, too, in the question of the freedom of the will it is important not to overshoot the scientific target, namely the explanation of physical and mental phenomena, in the direction of the unity of the physical world, and to take semantic matters seriously. Philosophy should take account of scientific developments, but science should also acknowledge philosophical distinctions. It appears that the conception of pragmatic dualism provides the best basis for this.

[20] Also the Libet experiments, according to which 300 milliseconds before a conscious 'act of the will' takes place, the corresponding readiness potential may already be measured, only to yield the desired conclusion, that we are not free in our decisions but determined by natural causalities if the muscle contraction taking place after the readiness potential has built up may be interpreted as an act of the will or expression of such an act. But precisely this needs to be also justified.

Neuroscience of Self-Consciousness and Subjectivity

Olaf Blanke[1]

Abstract

Recent data that have linked self-consciousness to the processing of multisensory bodily signals in human temporo-parietal and premotor cortex. Studies in which subjects receive ambiguous multisensory information about the location and appearance of their own body have shown that activity in these brain areas reflects the conscious experience of identifying with the body (self-identification), the experience of where "I" am in space (self-location), and the experience of the perspective from where "I" perceive the world (first-person perspective). I argue that these findings may form the basis for a neurobiological model of self-consciousness, grounding higher-order notions of self- consciousness and personhood in multisensory brain mechanisms.

Introduction

Human adults experience a "real me" that "resides" in "my" body and is the subject or "I" of experience and thought. This is self-consciousness, the feeling that conscious experiences are bound to the self. Thus, experiences are felt as belonging to "somebody" and it is this unitary entity, the "I", that is often considered to be one of the most astonishing features of the human mind.

A powerful approach to investigate self-consciousness has been to target brain mechanisms that process bodily signals (i.e. bodily self-consciousness).[1-6,54] The study of such bodily signals is complex as they are continuously present and updated and are conveyed by different senses as well as motor and visceral signals. Recent developments, however, in video, virtual reality and robotics technologies have allowed us to investigate the central mechanisms of bodily self-consciousness by providing subjects with ambiguous multisensory information about the location and appearance of

[1] Center for Neuroprosthetics, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; Laboratory of Cognitive Neuroscience, Brain Mind Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; Department of Neurology, University Hospital, 1211 Geneva, Switzerland.

their own body. The detailed study of abnormal states of bodily self-consciousness in neurological patients has also provided important additional data. Here I discuss three important aspects of bodily self-consciousness, how they relate to the processing of bodily signals, and which functional and neural mechanisms they may share: self-identification with the body (i.e. the experience of owning a body), self-location (i.e. the experience of where I am in space), and the first-person perspective (i.e. the experience from where I perceive the world).

I start with the breakdown of bodily self-consciousness and will describe the results of detailed observations in neurological patients with abnormal self-identification, self-location, and first-person perspective. Next, the major experimental paradigms, behavioural results, and neuroimaging findings will be described with the goal to develop a data-driven neurobiological model of self-consciousness and subjectivity.

When the self leaves the body: the out-of-body experience

If you ever – while lying in bed and about to fall asleep – suddenly had the distinct impression of floating up near the ceiling and looking back down at your body on the bed, then it is likely that you had an out-of-body experience (OBE). Here is a description of an OBE by Sylvan Muldoon, one of the first authors to describe his own OBEs (and those of others) in great detail: "I was floating in the very air, rigidly horizontal, a few feet above the bed [...] I was moving toward the ceiling, horizontal and powerless [...] I managed to turn around and there [...] was another 'me' lying quietly upon the bed" (from Muldoon & Carrington, *The projection of the astral body*, 1929).

OBEs may be considered a bizarre departure from normal human experience. Yet, they are more than a mere curiosity and of relevance for science and the humanities. They are a distortion of bodily self-consciousness and the study of this phenomenon has lead to insights into the bodily foundations of self-consciousness, and has impacted experimental research in cognitive neuroscience. OBEs are striking phenomena because they challenge our everyday experience of the spatial unity of self and body: they challenge our experience of a "real me" that "resides" in my body and is the subject or "I" of experience and thought.[8]

Yet, OBEs are not rare, have been reported since time immemorial, and have been estimated to occur in about 5% of the general population.[8] During an OBE, the subject has the subjective feeling of being awake and experiences the "self", or center of awareness, as being located outside the physical body, at a somewhat elevated level (i.e. abnormal self-location). It

is from this elevated extrapersonal location that the patient's body and the world are perceived (i.e. abnormal first-person perspective).[7-9] Most subjects experience to see their own body as lying on the ground or in bed, and the experience tends to be described as vivid and realistic. Thus, self-identification with a body, that is the sensation of owning a body, is experienced at the elevated, disembodied location and not at the location of the physical body (i.e. abnormal self-identification). What causes this disunity between self and body and the changes in self-identification, self-location, and our everyday body-centered first-person perspective?

The neurology of self-consciousness: the right temporo-parietal junction

OBEs of neurological origin have been reported in patients suffering from many different etiologies,[7-9] such as migraine,[10] epilepsy,[7,8,11] but also after focal electrical cortical stimulation,[12,13] general anesthesia,[14] typhoid fever,[15] and spinal cord damage.[16] OBEs due to focal brain damage have allowed further insights and have linked OBEs with the right and left temporo-parietal junction (TPJ),[8,17,18] in particular the posterior superior temporal gyrus,[8] angular gyrus,[12,18] and the supramarginal gyrus.[13,17] Outside the TPJ, damage has been found in the precuneus[13] and fronto-temporal cortex (Devinsky *et al.*, 1989). A recent lesion analysis study using voxel-based lesion symptom mapping in the to date largest sample of patients with OBEs due to focal brain damage, however, revealed a well-localized origin at the junction of the right angular gyrus with the posterior superior temporal gyrus[19].

Based on the frequent association of OBEs with visuo-somatosensory illusions, abnormal vestibular sensations,[20,21] and the role of the TPJ in multisensory integration,[22,23] it has been suggested that OBEs (and abnormal self-identification, self-location, and first-person perspective) occur due to disturbed multisensory integration of bodily signals in (peri)personal space (somatosensory, visual and proprioceptive signals) and extrapersonal space (visual and vestibular signals).[8,19] Translating these clinical data into the research laboratory, it has been investigated shown that multisensory conflicts of bodily signals can alter bodily self-consciousness by inducing altered states of self-identification, self-location, and the first-person perspective.

Video Ergo Sum: the induction of illusory self-location through visuo-tactile conflict

Extending paradigms from cognitive neuroscience and multisensory perception of the upper limb to paradigms targeting full-body representations

recent research using video, virtual reality and/or robotic devices in combinations with neuroimaging has allowed to study self- consciousness. Here, I will describe experimental procedures that have been developed to induce changes in self-location, self-identification and first-person perspective in healthy subjects.[24-26] These experiments induce full-body-illusions or out-of-body-illusions and exploit visuo–tactile and visuo–vestibular conflicts. In most such research paradigms, a tactile stroking stimulus is repeatedly applied to the back or chest of a participant who is being filmed and simultaneously views (through a head-mounted display (HMD)) the stroking of a human body in a real-time film or virtual-reality animation. The video camera was placed 2 m behind the person, filming the participant's back from behind. Thus, participants viewed a video image of their body (the "virtual body") from an "outside", third-person perspective[26] while an experimenter stroked their back with a stick. The stroking was thus felt by the participants on their back and also seen on the back of the virtual body. The HMD displayed the stroking of the virtual body either in real time or not (using an online video-delay or offline pre-recorded data), generating synchronous and asynchronous visuo-tactile stimulation, respectively.

Under these conditions subjects self-identified with the seen virtual body (*Video ergo sum*) and such illusory self-identification with the virtual body was stronger during synchronous than during asynchronous stroking conditions[26] (for a similar approach see[24]). Self-location was measured by passively displacing the body of the blindfolded subject after the stroking period and then asking her to walk back to the original position. As predicted, self-location was experienced at a position that was closer to the virtual body, as if subjects were located "in front" of the position where they had been standing during the experiment (or as if they were located "out-of-the-body").[26] Later work confirmed that self-location towards and self-identification with the virtual body are strongly and systematically influenced by different visuo-tactile conflicts and can also be achieved these changes in the supine position.[25,27]

These changes in bodily self-consciousness are also associated with an alteration how co-applied visual stimuli interfere with the perception of tactile stimuli.[28,29] Such visuo-tactile interference is a behavioural index of whether visual and tactile stimuli are functionally perceived to be in the same spatial location.[28,30-33] Applied during the abovementioned paradigms and during states of illusory self-identification and self-location, it was found that visual stimuli seen at a position that is 2 meters in front of the subject's back and tactile stimuli (that were applied on the subject's back) were functionally perceived to be in the same spatial location (see also[28,30-35]). These

data provide robust perceptual evidence (based on reaction times and accuracy rates) that self-identification and self-location with a virtual body alters how the brain perceives stimuli applied to the subject's own physical body. Moreover, self-identification and self-location are also associated with physiological (i.e. skin conductance response to a threat directed towards the virtual body)[24] and nociceptive changes (i.e. pain thresholds during the full body illusion)[36].

Multisensory brain mechanisms of self-identification and self-location

A comprehensive fMRI study[27] of a full-body illusion reported that self-identification with a virtual body is associated with activity in bilateral ventral premotor cortex, left posterior parietal cortex, and the left putamen. The activity in these three regions was enhanced by visuo-tactile stimulation. An EEG study[37] linked self-identification and self-location with a virtual body to activity in bilateral medial sensorimotor cortices and medial PMC. In this EEG study, self-identification and self-location with a virtual body induced by synchronous versus asynchronous visual-tactile stimulation of the real and the virtual body was associated with differential suppression of alpha band power (8–13 Hz) oscillations in bilateral medial sensorimotor regions and medial premotor cortex. Alpha band oscillations over central areas (that is, the mu rhythm) has been linked to sensorimotor processing[38] and mu rhythm suppression is thought to reflect increased cortical activation in sensorimotor and/or premotor cortices.[39] Another fMRI study[25] found that self-identification with a virtual body is associated with activation in the right middle-inferior temporal cortex (partially overlapping with the extrastriate body area (EBA)), a region that is like the premotor cortex involved in the multisensory processing of human bodies.[40-43]

The first-person perspective and visuo-vestibular processing

These former experimental procedures were able to induce changes in self-identification and self-location, but did not report changes in the first-person perspective that are a crucial aspect of bodily self-consciousness and prominently altered in an out-of-body experience. Recently changes in first-person perspective have also been achieved using fMRI and robotics, while participants were in a supine position and viewed a virtual body that was filmed from an elevated position.[25] Despite identical visuo-tactile stimulation, half of the participants experienced looking upward towards the virtual body (Up-group), and half experienced looking down on the virtual body (Down-group) and these perspectival changes were associated with consistent changes in self-location in both groups. In addition, subjective

reports of elevated self-location and sensations of flying, floating, rising, lightness, (that are common in out-of-body experiences) were frequent in the Down-group and rare in the Up-group.[25] These data show that self-location depends on visuo-tactile stimulation and on the experienced direction of the first-person perspective and that different multisensory mechanisms underlie self-location versus self-identification (the latter was not found to depend on the first-person perspective). These changes in self-location and the first-person perspective were reflected in activity at the TPJ bilaterally.[25] TPJ activity peaked in the posterior superior temporal gyri, differed between synchronous and asynchronous stroking conditions, and depended on the experienced direction of the first-person perspective.

Based on these findings it has been argued that the dependence of self-location on the first- person perspective may be caused by visuo-vestibular mechanisms (Blanke, 2012). Thus, in the study by Ionta *et al.*,[25] participants viewed a visual image on the HMD that contained a conflict between the visual gravitational cues of the virtual body and the vestibular gravitational cues experienced by the participant's physical body: the body that was shown in these experiments was presented in a direction that was incongruent with the direction of veridical gravity. This probably caused the observed differences in the experienced direction of the first-person perspective, with participants from the Up-group relying more strongly on vestibular cues from the physical body (indicating downward gravity directed towards the physical body) than on visual gravitational cues from the virtual body (indicating downward gravity directed away from the physical body), whereas participants from the Down-group show the opposite pattern. Indeed, past vestibular research has revealed prominent inter-individual differences in visuo-vestibular integration with some subjects relying more strongly on visual signals and others more on vestibular signals.[44–47] This is further corroborated by findings in subjects with OBEs and subjects with the so-called inversion illusion. Accordingly, these neurologically and experimentally-induced changes in the experienced direction of the first-person perspective are due to abnormal signal integration of otolithic vestibular cues and visual cues (Blanke, 2012).

Self-consciousness in humans and non-human primates?

Based on the involvement of human posterior parietal and premotor cortex in self- identification[27] and the known properties of bimodal visuo-tactile neurons in these regions in non-human primates, the observed changes in self-identification may be due to stroking-induced changes in the size and position of trunk-centred bimodal visuo-tactile neurons with

respect to the virtual body that is seen on the HMD (Blanke, 2012). In brief, the visual receptive fields of such bimodal neurons would be enlarged following visuo-tactile stroking, and would after stroking also encode the more distant position of the seen virtual body in peripersonal space.[48] Experimentally induced changes in illusory self-identification with a fake or virtual body via video-based virtual reality systems may therefore be associated with a stroking-induced enlargement or alteration of the brain's peripersonal space representation (visual receptive fields) of such bimodal parietal and/or premotor neurons. The described changes in self-identification may therefore be based on displaced or enlarged visual receptive fields of such bimodal neurons in premotor and posterior parietal cortex, so that they now also encode the more distant position of the seen virtual body. However, the reviewed neuroimaging and neurological data also suggest that there are differences between the brain mechanisms of self-location and the first-person perspective versus self-identification. Thus, self-location, but not self-identification, has been shown to depend on the first-person perspective, and relies on additional vestibular graviceptive (otolithic) signals and their integration with visual graviceptive signals. These distinct processes may recruit distinct brain regions in the posterior parietal cortex and the TPJ[49-51] (the parieto-insular vestibular cortex (PIVC), VIP and the middle superior temporal region (MST).[52]

Conclusions

The present data highlight the primary role of the temporo-parietal cortex in bodily self- consciousness as informed by multisensory and vestibular signals. Self-identification depends on somatosensory and visual signals and involves bimodal visuo-tactile neurons, whereas self-location and the first-person perspective depend on the integration of these bodily signals with vestibular cues, in trimodal visuo-tactile-vestibular neurons. These differences between self-identification versus self-location and first-person perspective are corroborated by neuroimaging and neurological data, showing that self-identification recruits primarily bilateral premotor and parietal regions, whereas self-location and the first-person perspective recruit posterior parietal–TPJ regions with a right hemispheric predominance.

These recent data extend other prominent proposals concerning the neural basis of bodily self-consciousness that have highlighted brain processes related to internal states of the body, such as interoceptive and homeostatic systems (e.g. the heartbeat) as important signals, and that have highlighted the contribution of either the insula[53] or the posterior medial parietal cortex.[54,55] Ongoing research explores the interactions between ex-

teroceptive bodily signals (which the present review focused on) and interoceptive and sensorimotor signals5,56. Recent results confirm that both types of bodily signals (extero- and interoceptive signals) are of relevance for self-consciousness and should despite their neuroanatomical differences be considered as a single system. These more recent findings also highlight the role of emotional mechanisms related to self-identification.[57]

Bodily self-consciousness may also turn out to be an important component for consciousness generally.[55] As Gerald Edelman stated "it is not enough to say that the mind (and consciousness) is embodied. You also have to say how".[58] Bodily self-consciousness may provide this link. Cognitive psychologists and neuroscientists have studied many different aspects of the self-related for example to language and memory (i.e. [2,4,5,59-68]). Along this line, mechanisms of bodily self-consciousness overlap with self-related processes such as perceptual and imagined viewpoint changes,[68] theory-of-mind, mentalizing,[69] and empathy.

The "I" of conscious experience is one of the most astonishing features of the human mind. The reviewed neuroscientific investigations of self-identification, self-location and first-person perspective have described some of the multisensory brain processes that may give rise to bodily self-consciousness. As argued elsewhere,[1] these three aspects are the necessary constituents of the simplest form of self-consciousness that arises when the brain encodes the origin of the first-person perspective from within a spatial frame of reference (i.e. self- location) associated with self-identification. It will be an exciting endeavour to better understand how the reviewed brain mechanisms are linked to language (i.e. [70,71]) and to memory and future prediction[55] (see also [72,73]) to also study higher-order – "narrative" and "extended" – aspects of self-consciousness and personhood.

Acknowledgements

Olaf Blanke is supported by the Bertarelli Foundation, the Swiss National Science Foundation, and the European Science Foundation. I am grateful to Christian Pfeiffer, Shahar Arzy, Bigna Lenggenhager, Tej Tadi, Christophe Lopez, Jane Aspell, Silvio Ionta, and Lukas Heydrich for many discussions and their valuable critiques.

References

[1] Blanke, O. & Metzinger, T. Full-body illusions and minimal phenomenal selfhood. *Trends Cogn Sci* 13, 7-13, doi:10.1016/j.tics.2008.10.003 (2009).

[2] Christoff, K., Cosmelli, D., Legrand, D. & Thompson, E. Specifying the self for cognitive neuroscience. *Trends Cogn Sci* 15, 104-112, doi:10.1016/j.tics.2011.01.001 (2011).

[3] de Vignemont, F. Embodiment, ownership and disownership. *Conscious Cogn* 20, 82-93, doi:10.1016/j.concog.2010.09.004 (2011).

[4] Jeannerod, M. The mechanism of self-recognition in humans. *Behav Brain Res* 142, 1-15 (2003).

[5] Knoblich, G. Self-recognition: body and action. *Trends Cogn Sci* 6, 447-449 (2002).

[6] Legrand, D. Pre-reflective self-as-subject from experiential and empirical perspectives. *Conscious Cogn* 16, 583-599, doi:10.1016/j.concog.2007.04.002 (2007).

[7] Devinsky, O., Feldmann, E., Burrowes, K. & Bromfield, E. Autoscopic phenomena with seizures. *Arch Neurol* 46, 1080-1088 (1989).

[8] Blanke, O., Landis, T., Spinelli, L. & Seeck, M. Out-of-body experience and autoscopy of neurological origin. *Brain* 127, 243-258, doi:10.1093/brain/awh040 [pii] (2004).

[9] Brugger, P. Reflective mirrors: perspective-taking in autoscopic phenomena. *Cogn Neuropsychiatry* 7, 179-194, doi:8H2GK5U13KJB8C20 [pii] 10.1080/13546800244000076 (2002).

[10] Lippman, C.W. Hallucinations of physical duality in migraine. *J Nerv Ment Dis* 117, 345-350 (1953).

[11] Heydrich, L., Lopez, C., Seeck, M. & Blanke, O. Partial and full own-body illusions of epileptic origin in a child with right temporoparietal epilepsy. *Epilepsy & Behavior* 20, 583-586 (2011).

[12] Blanke, O., Ortigue, S., Landis, T. & Seeck, M. Stimulating illusory own-body perceptions. *Nature* 419, 269-270, doi:10.1038/419269a [pii] (2002).

[13] De Ridder, D., Van Laere, K., Dupont, P., Menovsky, T. & Van de Heyning, P. Visualizing out-of-body experience in the brain. *N Engl J Med* 357, 1829-1833, doi:357/18/1829 [pii] 10.1056/NEJMoa070010 (2007).

[14] Lopez, U., Forster, A., Annoni, J.M., Habre, W. & Iselin-Chaves, I.A. Near-death experience in a boy undergoing uneventful elective surgery under general anesthesia. *Paediatr Anaesth* 16, 85-88, doi:PAN1607 [pii] 10.1111/j.1460-9592.2005.01607.x (2006).

[15] Menninger-Lerchenthal, E. *Der eigene Doppelgänger*. (Huber, 1946).

[16] Overney, L.S., Arzy, S. & Blanke, O. Deficient mental own-body imagery in a neurological patient with out-of-body experiences due to cannabis use. *Cortex* 45, 228-235, doi:S0010-9452(08)00131-7 [pii] 10.1016/j.cortex.2008.02.005 (2009).

[17] Maillard, L., Vignal, J.P., Anxionnat, R. & Taillandier Vespignani, L. Semiologic value of ictal autoscopy. *Epilepsia* 45, 391-394, doi:10.1111/j.0013-9580.2004.39103.x EPI39103 [pii] (2004).

[18] Brandt, C., Brechtelsbauer, D., Bien, C.G. & Reiners, K. [Out-of-body experience as possible seizure symptom in a patient with a right parietal lesion]. *Nervenarzt* 76, 1259, 1261-1252, doi:10.1007/s00115-005-1904-y (2005).

[19] Ionta, S. *et al.* Multisensory Mechanisms in Temporo-Parietal Cortex Support Self- Location and First-Person Perspective. *Neuron* 70, 363-374 (2011).

[20] Blanke, O. & Mohr, C. Out-of-body experience, heautoscopy, and autoscopic hallucination of neurological origin

Implications for neurocognitive mechanisms of corporeal awareness and self-consciousness. *Brain Res Brain Res Rev* 50, 184-199, doi:S0165-0173(05)00079-2 [pii] 10.1016/j.brainresrev.2005.05.008 (2005).

[21] Lopez, C., Halje, P. & Blanke, O. Body ownership and embodiment: vestibular and multisensory mechanisms. *Neurophysiol Clin* 38, 149-161, doi:S0987-7053(08)00018-X [pii] 10.1016/j.neucli.2007.12.006 (2008).

[22] Bremmer, F. *et al*. Polymodal motion processing in posterior parietal and premotor cortex: a human fMRI study strongly implies equivalencies between humans and monkeys. *Neuron* 29, 287-296, doi:S0896-6273(01)00198-2 [pii] (2001).

[23] Calvert, G.A., Campbell, R. & Brammer, M.J. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr Biol* 10, 649-657, doi:S0960-9822(00)00513-3 [pii] (2000).

[24] Ehrsson, H.H. The experimental induction of out-of-body experiences. *Science* 317, 1048, doi:10.1126/science.1142175 (2007).

[25] Ionta, S. *et al*. Multisensory mechanisms in temporo-parietal cortex support self-location and first-person perspective. *Neuron* 70, 363-374, doi:10.1016/j.neuron.2011.03.009 (2011).

[26] Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. *Video ergo sum*: manipulating bodily self-consciousness. *Science* 317, 1096-1099, doi:10.1126/science.1143439 (2007).

[27] Petkova, V.I. *et al*. From part- to whole-body ownership in the multisensory brain. *Curr Biol* 21, 1118-1122, doi:10.1016/j.cub.2011.05.022 (2011).

[28] Spence, C., Pavani, F. & Driver, J. Spatial constraints on visual-tactile cross-modal distractor congruency effects. *Cogn Affect Behav Neurosci* 4, 148-169 (2004).

[29] Aspell, J.E., Lenggenhager, B. & Blanke, O. Keeping in touch with one's self: multisensory mechanisms of self-consciousness. *PLoS One* 4, e6488, doi:10.1371/journal.pone.0006488 (2009).

[30] Igarashi, Y., Kimura, Y., Spence, C. & Ichihara, S. The selective effect of the image of a hand on visuotactile interactions as assessed by performance on the cross-modal congruency task. *Exp Brain Res* 184, 31-38, doi:10.1007/s00221-007-1076-z (2008).

[31] Pavani, F. & Castiello, U. Binding personal and extrapersonal space through body shadows. *Nat Neurosci* 7, 14-16, doi:10.1038/nn1167 (2004).

[32] Pavani, F., Spence, C. & Driver, J. Visual capture of touch: out-of-the-body experiences with rubber gloves. *Psychol Sci* 11, 353-359 (2000).

[33] Shore, D.I., Barnes, M.E. & Spence, C. Temporal aspects of the visuotactile congruency effect. *Neurosci Lett* 392, 96-100, doi:10.1016/j.neulet.2005.09.001 (2006).

[34] Aspell, J.E., Lavanchy, T., Lenggenhager, B. & Blanke, O. Seeing the body modulates audiotactile integration. *Eur J Neurosci* 31, 1868-1873, doi:10.1111/j.1460-9568.2010.07210.x (2010).

[35] Zopf, R., Savage, G. & Williams, M.A. Crossmodal congruency measures of lateral distance effects on the rubber hand illusion. *Neuropsychologia* 48, 713-725, doi:10.1016/j.neuropsychologia.2009.10.028 (2010).

[36] Hansel, A., Lenggenhager, B., von Kanel, R., Curatolo, M. & Blanke, O. Seeing and identifying with a virtual body decreases pain perception. *Eur J Pain* 15, 874-879, doi:10.1016/j.ejpain.2011.03.013 (2011).

[37] Lenggenhager, B., Halje, P. & Blanke, O. Alpha band oscillations correlate with illusory self-location induced by virtual reality. *Eur J Neurosci* 33, 1935-1943,

doi:10.1111/j.1460-9568.2011.07647.x (2011).
[38] Pineda, J.A. The functional significance of mu rhythms: translating "seeing" and "hearing" into "doing". *Brain Res Rev* 50, 57-68, doi:10.1016/j.brainresrev.2005.04.005 (2005).
[39] Oakes, T.R. *et al.* Functional coupling of simultaneous electrical and metabolic activity in the human brain. *Hum Brain Mapp* 21, 257-270, doi:10.1002/hbm.20004 (2004).
[40] Astafiev, S.V., Stanley, C.M., Shulman, G.L. & Corbetta, M. Extrastriate body area in human occipital cortex responds to the performance of motor actions. *Nat Neurosci* 7, 542-548, doi:10.1038/nn1241 (2004).
[41] Downing, P.E., Jiang, Y., Shuman, M. & Kanwisher, N. A cortical area selective for visual processing of the human body. *Science* 293, 2470-2473, doi:10.1126/science.1063414 (2001).
[42] Grossman, E.D. & Blake, R. Brain Areas Active during Visual Perception of Biological Motion. *Neuron* 35, 1167-1175 (2002).
[43] Urgesi, C., Candidi, M., Ionta, S. & Aglioti, S.M. Representation of body identity and body actions in extrastriate body area and ventral premotor cortex. *Nat Neurosci* 10, 30-31, doi:10.1038/nn1815 (2007).
[44] Golomer, E., Cremieux, J., Dupui, P., Isableu, B. & Ohlmann, T. Visual contribution to self-induced body sway frequencies and visual perception of male professional dancers. *Neurosci Lett* 267, 189-192 (1999).
[45] Isableu, B., Ohlmann, T., Cremieux, J. & Amblard, B. Selection of spatial frame of reference and postural control variability. *Exp Brain Res* 114, 584-589 (1997).
[46] Lopez, C., Lacour, M., Magnan, J. & Borel, L. Visual field dependence-independence before and after unilateral vestibular loss. *Neuroreport* 17, 797-803, doi:10.1097/01.wnr.0000221843.58373.c8 (2006).
[47] Young, L.R., Oman, C.M., Watt, D.G., Money, K.E. & Lichtenberg, B.K. Spatial orientation in weightlessness and readaptation to earth's gravity. *Science* 225, 205-208 (1984).
[48] Maravita, A. & Iriki, A. Tools for the body (schema). *Trends Cogn Sci* 8, 79-86, doi:10.1016/j.tics.2003.12.008 (2004).
[49] Grusser, O.J., Pause, M. & Schreiter, U. Vestibular neurones in the parieto-insular cortex of monkeys (Macaca fascicularis): visual and neck receptor responses. *J Physiol* 430, 559-583 (1990).
[50] Guldin, W.O., Akbarian, S. & Grusser, O.J. Cortico-cortical connections and cytoarchitectonics of the primate vestibular cortex: a study in squirrel monkeys (Saimiri sciureus). *J Comp Neurol* 326, 375-401, doi:10.1002/cne.903260306 (1992).
[51] Guldin, W.O. & Grusser, O.J. Is there a vestibular cortex? *Trends Neurosci* 21, 254-259 (1998).
[52] Lopez, C. & Blanke, O. The thalamocortical vestibular system in animals and humans. *Brain Res Rev* 67, 119-146, doi:10.1016/j.brainresrev.2010.12.002 (2011).
[53] Craig, A.D. How do you feel – now? The anterior insula and human awareness. *Nat Rev Neurosci* 10, 59-70, doi:10.1038/nrn2555 (2009).
[54] Damasio, A. & Meyer, D.E. in *The neurology of consciousness* (eds S. Laureys & G. Tononi) 3-14 (Elsevier, 2009).
[55] Damasio, A.R. *The feeling of what happens: Body and emotion in the making of consciousness*, (Harcourt Brace, 1999).
[56] Pacherie, E. The phenomenology of action: a conceptual framework. *Cognition* 107, 179-217, doi:10.1016/j.cognition.2007.09.003 (2008).
[57] Tsakiris, M., Tajadura-Jimenez, A. &

Costantini, M. Just a heartbeat away from one's body: interoceptive sensitivity predicts malleability of body-representations. *Proc Biol Sci* 278, 2470-2476, doi:10.1098/rspb.2010.2547 (2011).

[58] Edelman, G. *Bright air, brilliant fire* (Basic Books, 1992).

[59] Arzy, S., Thut, G., Mohr, C., Michel, C.M. & Blanke, O. Neural basis of embodiment: distinct contributions of temporoparietal junction and extrastriate body area. *J Neurosci* 26, 8074-8081, doi:10.1523/JNEUROSCI.0745-06.2006 (2006).

[60] Esslen, M., Metzler, S., Pascual-Marqui, R. & Jancke, L. Pre-reflective and reflective self-reference: a spatiotemporal EEG analysis. *Neuroimage* 42, 437-449, doi:10.1016/j.neuroimage.2008.01.060 (2008).

[61] Gillihan, S.J. & Farah, M.J. Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychol Bull* 131, 76-97, doi:10.1037/0033-2909.131.1.76 (2005).

[62] Heatherton, T.F. *et al.* Medial prefrontal activity differentiates self from close others. *Soc Cogn Affect Neurosci* 1, 18-25, doi:10.1093/scan/nsl001 (2006).

[63] Legrand, D. & Ruby, P. What is self-specific? Theoretical investigation and critical review of neuroimaging results. *Psychol Rev* 116, 252-282, doi:10.1037/a0014172 (2009).

[64] Macrae, C.N., Moran, J.M., Heatherton, T.F., Banfield, J.F. & Kelley, W.M. Medial prefrontal activity predicts memory for self. *Cereb Cortex* 14, 647-654, doi:10.1093/cercor/bhh025 (2004).

[65] Northoff, G. *et al.* Self-referential processing in our brain – a meta-analysis of imaging studies on the self. *Neuroimage* 31, 440-457, doi:10.1016/j.neuroimage.2005. 12.002 (2006).

[66] Perrin, F. *et al.* Neural mechanisms involved in the detection of our first name: a combined ERPs and PET study. *Neuropsychologia* 43, 12-19, doi:10.1016/j.neuropsychologia.2004.07.002 (2005).

[67] Platek, S.M. *et al.* Neural substrates for functionally discriminating self-face from personally familiar faces. *Hum Brain Mapp* 27, 91-98, doi:10.1002/hbm.20168 (2006).

[68] Vogeley, K. & Fink, G.R. Neural correlates of the first-person-perspective. *Trends Cogn Sci* 7, 38-42 (2003).

[69] Saxe, R., Moran, J.M., Scholz, J. & Gabrieli, J. Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Soc Cogn Affect Neurosci* 1, 229-234, doi:10.1093/scan/nsl034 (2006).

[70] Dennett, D. C. *Consciousness explained.*, (Penguin Books, 1991).

[71] Gazzaniga, M.S., LeDoux, J.E. & Wilson, D.H. Language, praxis, and the right hemisphere: clues to some mechanisms of consciousness. *Neurology* 27, 1144-1147 (1977).

[72] Arzy, S., Arzouan, Y., Adi-Japha, E., Solomon, S. & Blanke, O. The 'intrinsic' system in the human cortex and self-projection: a data driven analysis. *Neuroreport* 21, 569-574 (2010).

[73] Arzy, S., Bick, A. & Blanke, O. Mental time in amnesia: evidence from bilateral medial temporal damage before and after recovery. *Cogn Neuropsychol* 26, 503-510, doi:10.1080/02643290903439178 (2009).

Towards a Neuroscientific Understanding of Free Will

Neural Mechanisms Underlying Human Choice in the Frontal Cortex

Tim Behrens[1]

Introduction

In order to study how animals make choices, a good place to start is to consider *why* animals make choices. For the moment, discard your preconceptions of choice that come with visions of agonised Shakespearean characters making decisions of mortal consequence. Instead consider a much simpler example – the choice to do anything at all, rather than the alternative of doing nothing. This simple consideration leads us to a central theme in all decision sciences. We act because there is *value* in acting. It prevents starvation, thirst, and predation, and promotes procreation. Because different courses of action will lead to outcomes of different *values,* the decision-making problem, at its simplest, is to "often select courses of action with high value, and seldom those with low value".

In neuroscience experiments, the value of different decisions is often controlled experimentally. Different options may, for example, lead to different monetary outcomes in human experiments, or different quantities or qualities of food or drink in animal experiments. When subjects are asked to decide between such options, many brain regions, and even single cells within these brain regions, show activity that represents the worth or *value* of different options to an individual. High and low value options will lead to different activity patterns in a cell or cell population. This type of neural activity is notable, because it is neither a re-representation of a sensory stimulus – the brain's "input" – nor is it an "output" such as a motor command that will elicit a particular action. Instead, it is a signature of internal computations that are related to the choice itself. For example, neurons in the orbitofrontal cortex (OFC) will signal the amount of food that a choice will result in, but only when the animal is hungry (Rolls et al., 1989). FMRI signals recorded in the neighbouring ventromedial prefrontal cortex (vmPFC) will signal strongly when people are offered chocolate bars, but not if the subjects are dieting (Hare et al., 2011). Cells in the Anterior Cingulate Cortex that will increase their activity at the prospect of a rewarding

[1] FMRIB Centre, University of Oxford, University College, London.

drink, will also suppress their activity if the animal must exert effort in order to get that drink because the required effort reduces the overall worth to the animal (Kennerley et al., 2011).

Such value-related brain signals can be found in single neuron and fMRI responses across much of the brain (Rushworth and Behrens, 2008; Serences, 2008; Wallis and Kennerley, 2010). A major challenge, therefore, is to understand what different contributions these different brain regions might make to valuation, choice and behavioural control. In this essay, I will focus on three neighbouring cortical regions along the medial wall of the frontal cortex: The anterior cingulate cortex (ACC), the ventromedial prefrontal cortex (vmPFC), and the rostral dorsomedial prefrontal cortex (dmPFC) (figure 1, p. 317). These three regions are of particular interest, partly because they demonstrate a paradigmatic example of how computational ideas can provide important insights into the different roles of apparently similar neural activity; but also because, despite their proximity in the brain, these three regions likely first appeared at very different points during mammalian evolution. By considering their different contributions to behavioural control, we can perhaps gain some insight into the evolution of the complex behaviours exhibited everyday by humans. The ACC is present in all mammals that have been studied; the granular layer iv in vmPFC makes this region likely to be an evolutionary adaptation specific to primates (Mackey and Petrides, 2010; Tsujimoto et al., 2011); it was thought possible that the dmPFC was a specialisation unique to humans or at least great apes, until the recent discovery of a possible homologous area in the macaque monkey (Sallet et al., 2011).

Behavioural adaptation and the Anterior Cingulate Cortex

Assuming that you have overcome the first hurdle, and are committed to doing something rather than nothing, you now face a new and very difficult problem: What should you do? Solving this task optimally is not only difficult; it is impossible. The array of actions you can take is infinite, and you are continually faced with this challenge at every moment in time. In order to constrain the problem, you need a mechanism that performs two key functions. First, it must tell you when is a good time to make a decision. I hope for example, that you will not decide to make a cup of tea between reading the next two words in this sentence. Second it must be able to reduce the candidate options to a number that might reasonably be compared.

One simple strategy that can help with such constraints is to give default preference to a particular course of action. Such an action might simply be the action that you are currently taking (reading this sentence), or might, for

example, be a habit that you have often followed in the same situation (Daw et al., 2005), such as going to work by the same route at the same time each morning. This strategy simplifies the decision-making problem dramatically. The problem is now one of knowing when to change your default strategy and adopt a new one. A wealth of evidence is suggestive that the dorsal ACC subserves computations that are tuned for exactly such a decision.

The ACC responds to changes in the world

In order to discern whether you should continue with your current policy, it is critical that you monitor the outcomes of your actions. Actions that often result in good outcomes should be repeated. Those that do not should be discarded. Neural responses in the ACC are tuned to the outcomes of actions, whether positive or negative (Jocham et al., 2009; Kennerley et al., 2011; Matsumoto et al., 2007; Shima and Tanji, 1998; Walton et al., 2004), and these responses are particularly strong if the outcome was a consequence of an action voluntarily selected by the subject (Walton et al., 2004). The ACC therefore has the opportunity to evaluate the quality of our current behavioural policy. To evaluate whether, and how well, it takes this opportunity, we must turn to some more mathematical ideas.

One reason why it might be important to change from your current preferred strategy is that something about the world has changed to mean that this strategy is no longer good. Your favourite apple tree has no apples, or your prey has moved away from the valley. But how do you know if a real change has occurred, or if your latest observation is just an aberration? The buffalo may have started their annual migration, or may have moved just for one day. This problem is equivalent to asking "How much should I learn from this latest piece of data in comparison to the rest of my experiences", and is a problem that Bayesian statistics is ideally placed to answer (Courville et al., 2006). A core idea in Bayesian statistics is that different pieces of information should be reconciled according to their respective predictive values. Hence in situations in which the most recent piece of information is a better predictor of the future than historical information (such as fast changing, or "volatile" environments) subjects should also learn at a fast rate by placing a great deal of weight on each new observation. By contrast, in situations where historical information is still informative (because the world changes only slowly), subjects should place little weight on new data points and instead, stay with their original policies.

Humans (Behrens et al., 2007; Nassar et al., 2010), macaques (Rushworth and Behrens, 2008) and rodents (Gallistel et al., 2001) do indeed learn faster when the environment is more volatile and, in Humans, each new obser-

vation causes more ACC activity in a volatile environment than that same observation would in a stable one (Behrens et al., 2007). Furthermore, individual subjects who display this ACC activity to a greater extent are the same subjects who are fastest to change their beliefs (Behrens et al., 2007). The anterior cingulate cortex is not simply monitoring the outcomes of our choices, but also using these outcomes to optimally adapt our future behaviour (Rushworth and Behrens, 2008). Indeed, if lesions are made to the ACC in macaque monkeys the impairment that can be observed is precisely this capacity to integrate past observations appropriately to guide future behavioural change (Kennerley et al., 2006).

These data reveal insights about the types of computations that occur in the ACC to guide choices, but more recently there have also been insights into how these computations occur at the level of cellular dynamics and neurochemicals. When Rats perform the simple task of choosing the correct location to find food, they are able to maintain a stable strategy even in the face of noisy outcomes. If the food is in the same place on 70 or 80% of occasions this will be the place they look first, even if it was somewhere else last time round. During these periods of stable belief, the pattern of activity amongst a population of cells in the ACC also remains stable (Karlsson et al., 2012) - each cell's activity will look the same on this trial as it did on the last. If the experimenter then plays a trick on the animal, and changes the best location to find the food, after several errors the rat will eventually learn to change strategy. However, before the rat commits to a new strategy, it undergoes a period of uncertainty when it is modifying its internal beliefs, displaying a decreased resolve to pursue a single strategy and, instead, exploring the different options seemingly at random. During this period, patterns of activity in ACC cells undergo volatile changes from trial to trial as the old belief is broken and the new one formed (Karlsson et al., 2012).

It is not clear what neural events lead to these rapid resets of the ACC cellular network, but one intriguing possibility is that they are mediated by the neuromodulator Norepinephrine (NE) (Yu and Dayan, 2005). The source of almost all of the brain's NE is the locus coeruleus (LC), a small nucleus in the mid-brain with a major input to the ACC. Release of NE from the locus coeruleus has broad ranging effects on the brain's arousal systems and, in particular, causes a dilation of the pupils. This convenient fact enables neuroscientists to measure an index of LC activity without harm to the subject. Matthew Nassar and colleagues have measured pupil diameter whilst subjects performed a change detection task (Nassar et al., 2012). In this task, there are two computational factors that should make

subjects amenable to changes of belief: The long-term probability that the world might change (akin to the environment volatility described above), and a term known as the relative uncertainty, which captures mathematically your doubt that your previous belief was correct. As these factors are varied by the experimenter throughout the experiment, they both exhibit strong and separable influences on the measured pupil diameter. Perhaps most impressively, if the experimenter introduces a surprising stimulus (a loud noise) at an unexpected time in the experiment, this not only causes an increase in pupil diameter, but also results in a rapid period of revising beliefs about the subject's completely unrelated task.

This causal intervention suggests that computations in the ACC are not simply responsible for detecting changes in the environment, but also for inducing resultant changes in behaviour. Such a computation is central to ecological theories of foraging animals(Charnov, 1976). If an animal is foraging for food in a field, the amount of food in that field will decline, but it is costly to leave the field and find a better one. If the animal's strategy is always to move to the best field, he will spend almost all of his time walking between fields! So when should he move from one field to the next? The mathematics of this problem can be solved, and there is indeed an optimal "foraging time" in each field(Charnov, 1976). When monkeys are asked to solve this problem in a laboratory they solve the problem almost exactly perfectly. They find the optimal foraging times no matter how rich is their own field, how rich is the competing field or how much it costs to travel between fields (Hayden et al., 2011). It seems that this remarkable capacity is due to this same ACC behavioural adaptation mechanisms. Whilst the animal is happily foraging in his patch, cells in the ACC are signalling exactly how good it would be to leave this patch right now. When these cells reach a threshold level of firing, the animal moves to the next patch (Hayden et al., 2011).

Two recent studies have demonstrated that these ancient foraging mechanisms are also at work when humans make decisions, and that they again rely on the anterior cingulate cortex. When humans are asked whether they would like to stay with a current option or to return to the world to see what they will get, FMRI signal in the ACC reflects the average value of everything else in the environment, and inversely reflects the cost of returning to the environment (Kolling et al., 2012). Just like the cells in Hayden's monkeys, signals in the human ACC reflect the expected value of changing from the current behaviour (Kolling et al., 2012). Furthermore, this signal works at the strategic level – exactly as is needed to solve the foraging problem. If subjects temporarily break from their long-term strat-

egy to try something new, the adaptation signal in the ACC both influences and later learns about the long-term strategy, and not the short term distraction (Boorman et al., in press).

Evaluative choices and the ventro-medial prefrontal cortex

Whilst many decision-making problems might be solved by the kinds of simple behavioural adaptation strategies such as those that I have attributed to the ACC, humans are certainly capable of choices that cannot be solved in such a simple fashion. We can decide to get married, to spend hundreds of thousands of pounds on a new house, or even make simple choices between restaurants where we havenever previously eaten. Two experiments that I have already described in the context of foraging-style choices above (Boorman et al., in press; Kolling et al., 2012) also compared these foraging choices with a particular type of 'evaluative choice' that can be examined in a laboratory. In both cases, whilst foraging-type activity was recorded in ACC, activity that reflected these evaluative choices could be recorded in the vmPFC (Boorman et al., in press; Kolling et al., 2012). That the vmPFC is particularly important for these types of choices has long been known from experiments that study patients with damage to vmPFC. Patients with vmPFC damage become indecisive about even trivial decisions (Barrash et al., 2000); choices that are made are often made poorly (Bechara et al., 1994; Bechara et al., 2000) according to unusual strategies (Fellows, 2006).

It is tempting to think that such subjective and complex behaviours as these might be immune to computational descriptions, but some progress has been made. Much like in the ACC, neural signals in vmPFC encode the value of potential choices at both the single cell (Bouret and Richmond, 2010)((Monosov and Hikosaka, 2012) and population level (FitzGerald et al., 2009; Kable and Glimcher, 2009; Rangel and Hare, 2010). However, vmPFC responses appear particularly flexible. Whilst many other brain regions rely on direct experience of previous outcomes to estimate the value of different courses of action, vmPFC can encode values that must be computed on the fly. These computations may, for example, rely on an understanding of the complex structure of the environment (Hampton et al., 2006); from the generalisation of concepts learnt in different situations (Kumaran et al., 2009); or from the integration of several disparate sources of information (Behrens et al., 2008). Perhaps most strikingly, if subjects are asked to ignore all of their own experiences and preferences, and instead to guess what a very different individual would choose, vmPFC value signals immediately reflect the preferences of this new individual (Janowski et al., 2012; Nicolle et al., 2012). If,

however, the problem at hand is best solved by considering values learnt from direct experience, the vmPFC can seamlessly revert to these more basic value computations (Wunderlich et al., 2012).

We are only now beginning to investigate the mechanisms that allow vmPFC and connected brain regions to perform these complex evaluations (Wimmer et al., 2012), but more progress has been made in understanding how a network of cells might use these computed values to select choices and guide behaviour. Such explicit or evaluative choices do not benefit from the simplifying strategies employed in foraging-style choices (Boorman et al., in press), so this system is once again faced with the problem of focussing attention on the option with the highest expected value from many possible alternatives. One possible solution to this problem is for different options to compete simultaneously for neural representation. Particular patterns of cellular activity associated with each option may become excited if that option is potentially valuable, and this activity may inhibit other representations from forming (e.g. (Wang, 2002)). Such a neural architecture implies a competition that is seamlessly resolved by options inhibiting each other until activity only remains in a single one. At this point, a decision has been made. If the competing neural representations are initially more excited by more valuable potential outcomes, then such architecture will, on average, make profitable decisions.

It is possible to construct neural networks with exactly this architecture in silico (e.g. (Wang, 2002, 2008)) and examine how they behave when faced with the same choices made by laboratory subjects. When the activity in such a simulated network is analysed, a complex and precise pattern can be seen in the average activity of the network that is a signature of this competitive inhibitory architecture. The activity of the network transforms its representation of value midway through the decision, does so with particular timings and at particular frequencies. When we look for a region that expresses this signature in the human brain, we find this exact pattern of activity in the vmPFC (Hunt et al., 2012). It appears that a competition mechanism exists in vmPFC that can highlight favourable options and suppress unfavourable or irrelevant ones. Indeed, when lesions are made to macaque vmPFC, unfavourable distracting options cannot be suppressed and interfere with the animal's choices (Noonan et al., 2010).

Again, however, modern methods allow us to go further and examine the neurochemistry that underlies these cortical computations. If it is indeed the case that competitions can be resolved by combining the excitation of favourable options with the inhibition of irrelevant ones then these competitions likely rely on the brain's major excitatory and inhibitory neuro-

transmitters, Glutamate and GABA. Indeed, using the same *in-silico* architectures that predicted dynamic signatures of choice above, one can predict exactly *how* variations in GABA and Glutamate concentrations should affect both subject choices and neural dynamics (Jocham et al., 2012). Increasing the GABAergic inhibition in such simulated networks results in choices that are resolved more slowly, but more accurately; increasing the glutamatergic excitation results in faster more erratic choices. Combined with the knowledge that cortical Glutamate and GABA concentrations may vary substantially across different individuals (Stagg et al., 2011), these observations make a rather surprising prediction: a behaviour as complex and personal as value-guided choice might depend predictably on basic neurochemistry. And indeed it does - People with high GABA concentrations in vmPFC exhibit slow neural dynamics and accurate decisions. Those with high vmPFC Glutamate concentrations exhibit fast dynamics and erratic decisions (Jocham et al., 2012).

Modelled choice and the dorsomedial prefrontal cortex

Sitting just above the vmPFC on the medial surface of the prefrontal cortex (figure 1, p. 317) is a cortical area that is much less well studied and understood. The dmPFC is a region of cortex in which no single cell activity has ever been recorded because, until recently (Sallet et al., 2011), it has not been clear that the region even exists in any nonhuman species. In humans, the signals that can be recorded in this region are extremely alluring. The region, for example, appears to be particularly active in situations in which the experimental subject must attribute motive or intention to something in their surroundings. For example, dmPFC activity is weak when a subject is viewing triangles moving around a screen, but strong if one triangle appears to coax another to move in a particular direction (Castelli et al., 2002; Castelli et al., 2000). Responses such as these have led scientists to suggest that the region might play an important role in *social* cognition. The ability to infer the intentions and likely actions of others is of clear evolutionary value to all social animals, and humans are perhaps unique in the number and complexity of their social interactions (Dunbar and Shultz, 2007).

It is likely, then, that this brain region that has appeared relatively recently in human evolution might support a type of activity that underlies this complex human behaviour, but can we discover how? It seems to be a daunting challenge. Is it even possible to describe such mechanisms in a way that is amenable to scientific testing? It is very early days, but there are some indications that this problem might not be completely beyond our reach. One approach that has been taken is to look for parallels between mechanisms that

might underlie social behaviour, and those that are known about in non-social settings. We know, for example, that a key neural mechanism that controls reward learning is computation of the *reward prediction error* – the difference in reward between what was expected and what was received (Schultz et al., 1997). Neural signals that code for the reward prediction error are most famously found in the dopaminergic ventral tegmental area (Schultz et al., 1997) but can also be found in other brain regions including the vmPFC (Rutledge et al., 2010). Might a similar mechanism underlie our inferences about other individuals? Indeed, when subjects are asked to learn about the likely intentions of a confederate player in a game, FMRI activity in dmPFC first reflects first a prediction, and then a prediction error on the confederate's actions (Behrens et al., 2008; Hampton et al., 2008). Notably, in these studies the predictions and prediction errors did not concern the value of actions, but instead the truth of communicative intentions.

That the mechanisms at play in dmPFC (and related social regions (Frith and Frith, 2012)) might parallel those in vmPFC raises questions about whether there is really anything unique about the social nature of these computations. There is another intriguing possibility. By their very nature, representations of other individuals almost always require a model of that individual that is separate from the subject's current sensory and motor environment. In order to predict what another individual will do, the relevant environment is theirs, not yours. It is possible that it is this capacity to perform processing that is abstracted from the senses that was the key contribution brought by the evolution of dmPFC (Frith and Frith, 2003). If so, it is possible that the mechanisms and computations in dmPFC have a general similarity to those in vmPFC but take place in this abstracted frame of reference. Once abstracted, the values and goals that are represented in vmPFC activity might easily be misconstrued as dmPFC's motives and intentions.

It is possible to test such an idea by constructing an experiment similar to the simple value comparison task that was informative about vmPFC competition mechanisms, but with a twist. On some trials the subjects should choose not for themselves, but instead for another individual with very different preferences to their own. In these trials, then, the subject is choosing according to their own sensory and motor environment, but their partner's valuations (Nicolle et al., 2012). On trials when the subjects choose for themselves, signatures of their own choices can be seen vmPFC. In dmPFC, despite their irrelevance to the current task, the exact same signatures can be seen, but here computed according to the partner's values and the choices that the partner would have made. That it is the exact same signature that is recorded in the two regions supports the idea of similar un-

derlying computations. Crucially, however, when the subjects now choose on behalf of their partner, the two brain areas again exhibit the same signature but exchange agents so that the vmPFC now represents the partner's choices and the dmPFC the subject's. dmPFC activity is not required to model the other individual, but rather to abstract the choice from the immediate environment (Nicolle et al., 2012). Indeed, it is possible that the capacity to perform such abstract value processing has important functions that are completely divorced from social processing. For example, some decisions may rely on the ability to model ones own likely behavior in the context of future choices that ensue after the immediate action.

Conclusions

I have tried to demonstrate how modern scientists are attempting to dissect the neural mechanisms that control our behaviour. I have argued that humans display behaviours that are common to many other animals, and that in order to understand the neural processes that support these behaviours, we must first understand how the behaviours themselves evolved. Furthermore, I have argued that it is informative to consider the evolutionary precursors to human behaviours, even for behaviours (and perhaps neural processes) that are uniquely human. In making this argument, I have attempted to demonstrate several situations in which modern neuroscience has not only uncovered the computations that are being performed in different brain regions, but also how those computations might be performed in networks of cells and neurochemicals. Such an approach might even help us understand how the most complex human behaviours emerge - a major goal of neuroscience, and one that we are beginning to tackle.

References

Barrash, J., Tranel, D., Anderson, S.W., 2000. Acquired personality disturbances associated with bilateral damage to the ventromedial prefrontal region. Dev Neuropsychol 18, 355-381.

Bechara, A., Damasio, A.R., Damasio, H., Anderson, S.W., 1994. Insensitivity to future consequences following damage to human prefrontal cortex. Cognition 50, 7-15.

Bechara, A., Damasio, H., Damasio, A.R., 2000. Emotion, decision making and the orbitofrontal cortex. Cereb Cortex 10, 295-307.

Behrens, T.E., Hunt, L.T., Rushworth, M.F., 2009. The computation of social behavior. Science 324, 1160-1164.

Behrens, T.E., Hunt, L.T., Woolrich, M.W., Rushworth, M.F., 2008. Associative learning of social value. Nature 456, 245-249.

Behrens, T.E., Woolrich, M.W., Walton, M.E., Rushworth, M.F., 2007. Learning the value of information in an uncertain world. Nat Neurosci 10, 1214-1221.

Boorman, E.D., Rushworth, M.F., Behrens, T.E., in press. Ventromedial prefrontal and anterior cingulate cortex adopt choice

and default reference frames during multialternative sequential choice. Journal of Neuroscience.

Bouret, S., Richmond, B.J., 2010. Ventromedial and orbital prefrontal neurons differentially encode internally and externally driven motivational values in monkeys. J Neurosci 30, 8591-8601.

Castelli, F., Frith, C., Happe, F., Frith, U., 2002. Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. Brain 125, 1839-1849.

Castelli, F., Happe, F., Frith, U., Frith, C., 2000. Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. Neuroimage 12, 314-325.

Charnov, E.L., 1976. Optimal Foraging, The Margnial Value Theorem. Theoretical Population Biology 9.

Courville, A.C., Daw, N.D., Touretzky, D.S., 2006. Bayesian theories of conditioning in a changing world. Trends in cognitive sciences 10, 294-300.

Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature neuroscience 8, 1704-1711.

Dunbar, R.I., Shultz, S., 2007. Evolution in the social brain. Science 317, 1344-1347.

Fellows, L.K., 2006. Deciding how to decide: ventromedial frontal lobe damage affects information acquisition in multi-attribute decision making. Brain : a journal of neurology 129, 944-952.

FitzGerald, T.H., Seymour, B., Dolan, R.J., 2009. The role of human orbitofrontal cortex in value comparison for incommensurable objects. J Neurosci 29, 8388-8395.

Frith, C.D., Frith, U., 2012. Mechanisms of social cognition. Annu Rev Psychol 63, 287-313.

Frith, U., Frith, C.D., 2003. Development and neurophysiology of mentalizing. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 358, 459-473.

Gallistel, C.R., Mark, T.A., King, A.P., Latham, P.E., 2001. The rat approximates an ideal detector of changes in rates of reward:implications for the law of effect. J Exp Psychol Anim Behav Process 27, 354-372.

Hampton, A.N., Bossaerts, P., O'Doherty, J.P., 2006. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci 26, 8360-8367.

Hampton, A.N., Bossaerts, P., O'Doherty, J.P., 2008. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci U S A 105, 6741-6746.

Hare, T.A., Malmaud, J., Rangel, A., 2011. Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice. J Neurosci 31, 11077-11087.

Hayden, B.Y., Pearson, J.M., Platt, M.L., 2011. Neuronal basis of sequential foraging decisions in a patchy environment. Nat Neurosci 14, 933-939.

Hunt, L.T., Kolling, N., Soltani, A., Woolrich, M.W., Rushworth, M.F., Behrens, T.E., 2012. Mechanisms underlying cortical activity during value-guided choice. Nature neuroscience 15, 470-476, S471-473.

Janowski, V., Camerer, C., Rangel, A., 2012. Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. Soc Cogn Affect Neurosci.

Jocham, G., Hunt, L.T., Near, J., Behrens, T.E., 2012. A mechanism for value-guided choice based on the excitation-inhibition balance in prefrontal cortex. Nature neuroscience 15, 960-961.

Jocham, G., Neumann, J., Klein, T.A., Danielmeier, C., Ullsperger, M., 2009. Adaptive coding of action values in the

human rostral cingulate zone. J Neurosci 29, 7489-7496.
Kable, J.W., Glimcher, P.W., 2009. The neurobiology of decision: consensus and controversy. Neuron 63, 733-745.
Karlsson, M.P., Tervo, D.G., Karpova, A.Y., 2012. Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. Science 338, 135-139.
Kennerley, S.W., Behrens, T.E., Wallis, J.D., 2011. Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. Nat Neurosci.
Kennerley, S.W., Walton, M.E., Behrens, T.E., Buckley, M.J., Rushworth, M.F., 2006. Optimal decision making and the anterior cingulate cortex. Nat Neurosci 9, 940-947.
Kolling, N., Behrens, T.E., Mars, R.B., Rushworth, M.F., 2012. Neural mechanisms of foraging. Science 336, 95-98.
Kumaran, D., Summerfield, J.J., Hassabis, D., Maguire, E.A., 2009. Tracking the emergence of conceptual knowledge during human decision making. Neuron 63, 889-901.
Mackey, S., Petrides, M., 2010. Quantitative demonstration of comparable architectonic areas within the ventromedial and lateral orbital frontal cortex in the human and the macaque monkey brains. The European journal of neuroscience 32, 1940-1950.
Matsumoto, M., Matsumoto, K., Abe, H., Tanaka, K., 2007. Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci 10, 647-656.
Monosov, I.E., Hikosaka, O., 2012. Regionally distinct processing of rewards and punishments by the primate ventromedial prefrontal cortex. The Journal of neuroscience : the official journal of the Society for Neuroscience 32, 10318-10330.
Nassar, M.R., Rumsey, K.M., Wilson, R.C., Parikh, K., Heasly, B., Gold, J.I., 2012. Rational regulation of learning dynamics by pupil-linked arousal systems. Nature neuroscience 15, 1040-1046.
Nassar, M.R., Wilson, R.C., Heasly, B., Gold, J.I., 2010. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. The Journal of neuroscience : the official journal of the Society for Neuroscience 30, 12366-12378.
Nicolle, A., Klein-Flugge, M.C., Hunt, L.T., Vlaev, I., Dolan, R.J., Behrens, T.E., 2012. An agent independent axis for executed and modeled choice in medial prefrontal cortex. Neuron 75, 1114-1121.
Noonan, M.P., Walton, M.E., Behrens, T.E., Sallet, J., Buckley, M.J., Rushworth, M.F., 2010. Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. Proc Natl Acad Sci U S A 107, 20547-20552.
Rangel, A., Hare, T., 2010. Neural computations associated with goal-directed choice. Curr Opin Neurobiol 20, 262-270.
Rolls, E.T., Sienkiewicz, Z.J., Yaxley, S., 1989. Hunger Modulates the Responses to Gustatory Stimuli of Single Neurons in the Caudolateral Orbitofrontal Cortex of the Macaque Monkey. Eur J Neurosci 1, 53-60.
Rudebeck, P.H., Walton, M.E., Smyth, A.N., Bannerman, D.M., Rushworth, M.F., 2006. Separate neural pathways process different decision costs. Nat Neurosci 9, 1161-1168.
Rushworth, M.F., Behrens, T.E., 2008. Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Neurosci 11, 389-397.
Rutledge, R.B., Dean, M., Caplin, A., Glimcher, P.W., 2010. Testing the reward prediction error hypothesis with an axiomatic model. The Journal of neuroscience : the official journal of the Society for Neuroscience 30, 13525-13536.
Sallet, J., Mars, R.B., Noonan, M.P., Andersson, J.L., O'Reilly, J.X., Jbabdi, S.,

Croxson, P.L., Jenkinson, M., Miller, K.L., Rushworth, M.F., 2011. Social network size affects neural circuits in macaques. Science 334, 697-700.

Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275, 1593-1599.

Serences, J.T., 2008. Value-based modulations in human visual cortex. Neuron 60, 1169-1181.

Shima, K., Tanji, J., 1998. Role for cingulate motor area cells in voluntary movement selection based on reward. Science 282, 1335-1338.

Stagg, C.J., Bachtiar, V., Johansen-Berg, H., 2011. The role of GABA in human motor learning. Current biology : CB 21, 480-484.

Tsujimoto, S., Genovesio, A., Wise, S.P., 2011. Frontal pole cortex: encoding ends at the end of the endbrain. Trends in cognitive sciences 15, 169-176.

Wallis, J.D., Kennerley, S.W., 2010. Heterogeneous reward signals in prefrontal cortex. Current opinion in neurobiology 20, 191-198.

Walton, M.E., Devlin, J.T., Rushworth, M.F., 2004. Interactions between decision making and performance monitoring within prefrontal cortex. Nat Neurosci 7, 1259-1265.

Wang, X.J., 2002. Probabilistic decision making by slow reverberation in cortical circuits. Neuron 36, 955-968.

Wang, X.J., 2008. Decision making in recurrent neuronal circuits. Neuron 60, 215-234.

Wimmer, G.E., Daw, N.D., Shohamy, D., 2012. Generalization of value in reinforcement learning by humans. The European journal of neuroscience 35, 1092-1104.

Wunderlich, K., Dayan, P., Dolan, R.J., 2012. Mapping value based planning and extensively trained choice in the human brain. Nature neuroscience 15, 786-791.

Yu, A.J., Dayan, P., 2005. Uncertainty, neuromodulation, and attention. Neuron 46, 681-692.

False Perceptions & False Beliefs: Understanding Schizophrenia

Chris D. Frith[1] & Karl J. Friston[2]

False perceptions and false beliefs are core symptoms of schizophrenia. If we want to understand these symptoms, we need to explore the interaction between the physical and the mental. Schizophrenia provides a unique opportunity to explore this interaction. To explain how the mental emerges from the physical is the key challenge facing 21st century science.

1. The symptoms

Hallucinations (false perceptions) and delusions (false beliefs) are characteristic symptoms of schizophrenia. Typical hallucinations include: *hearing people talking to you or about you, hearing a running commentary on your actions,* and *hearing your thoughts spoken aloud*. Typical delusions include; *believing that other people can hear your thoughts, believing that your actions are being controlled by external forces,* and *believing that people are sending you secret messages* (Mellor 1970). When reporting these symptoms, patients are trying to describe extremely unusual experiences and, as is indicated by the typical verbatim examples given below, the symptom labels listed above do not fully capture these experiences (examples from Kambeitz-Ilankovic *et al.* 2012).

> *If I breathe without other people then they get stuck to me. I get stuck to people and the thoughts come through people. There are things I've learned just before I came in. It was so bad I could hear everybody in my mind. It is like being stuck on the same wavelength as people.*
>
> *I felt myself touched in such a way as if I were hypnotised, electrified, or generally controlled by some sort of medium or some other will.*

2. The problem

These strange and frightening experiences lie in the mental domain, and may occur in the absence of any obvious changes in behaviour. However, these subjective experiences are intimately connected with physical events in the

[1] Wellcome Trust Centre for Neuroimaging at University College, London; Interacting Minds Centre, Aarhus University; All Souls College, Oxford.
[2] All Souls College, Oxford.

brain. It has been known for 50 years that the neurotransmitter dopamine has a role in the generation of the symptoms of schizophrenia. Treatment with drugs that block dopamine receptors reduce the severity of hallucinations and delusions (Johnstone *et al.* 1978). In contrast, drugs, such as amphetamine that increase the levels of dopamine in the brain, can cause symptoms very similar to those associated with schizophrenia (Connell 1958). More recently, advanced imaging methods have demonstrated increased dopamine activity in the mid–brain (striatum) of people with schizophrenia. This excessive activity is probably present in the prodromal phase of the illness before the appearance of florid psychotic symptoms (Frith 2002).

Our approach has been to develop a cognitive account of the particular symptoms associated with schizophrenia. The use of a cognitive framework, that is a computational approach based on cybernetics and information theory, is very useful since terms, such as information and representation, can be applied at the physiological as well as the psychological level of description. A successful cognitive account of particular symptoms should help us, first, to understand, a bit better, what the experience is like, second, to generalize the account to explain the whole range of hallucinations and delusions associated with schizophrenia, and third, to generalize further to how perceptions and beliefs are acquired in the normal case.

3. Explaining delusions of control

One of the more striking experiences reported by patients is labelled *delusion of control*. The patient feels that his actions are controlled by external forces (examples from Mellor 1970).

My fingers pick up the pen, but I don't control them. What they do is nothing to do with me.

The force moved my lips. I began to speak.

In common with a number of other symptoms (such as *hearing one's one thoughts spoken aloud*), this experience seems to spring from confusion between something that I do and something that is happening independently from me in the outside world. It has long been recognised that this distinction creates a problem for the nervous system (Helmholtz 1866). For example, when an image moves across my retina, how do I know whether this is because I am moving my eye, or because an object is moving past me? For the nervous system the difference between these two situations is that, when I move my eye, commands have been sent to the eye muscles to cause the movement to occur. Such commands have not been sent when the object moves past my eye.

Helmholtz's resolution of this problem (and many other aspects of perception) was to regard the brain as an inference machine – generating predictions about the sensory consequences of action (known as corollary discharge). Put simply, if I believe I am moving my eyes, then I will predict and confirm my re-sampling of the visual field. Conversely, if visual input changes in the absence of unintended eye movement, then the best hypothesis – that could explain this sensory evidence – is that the world is moving. Treating perception as hypothesis testing or (unconscious) inference is central to the arguments that follow and is particularly important for attribution of agency: inferring the causes of changing sensory input requires a judicious balance between the precision or confidence I assign my prior beliefs (or hypotheses), relative to sensory evidence. If moving my eyes depends upon the prior belief that visual (and proprioceptive) signals will change, then assigning too much precision to the sensory consequences of moving will provide evidence against any movement, and will subvert the intended action. It is therefore necessary to attenuate the precision of sensory signals when, and only when, they report the consequences of intended movements. This is known as sensory attenuation, whereby sensations associated with voluntary movements are suppressed and ignored. This is why the sensations produced when we tickle ourselves are so much weaker than when someone else tickles us. On the other hand, if the expected feedback is manipulated and distorted then the intensity of the effects produced by our own movements is increased (Blakemore *et al.* 1999).

Several experiments have shown that patients, especially those with delusions of control, are abnormally aware of the sensations associated with voluntary movements; in other words, there is a failure to attenuate the precision of sensations. For example, they find the experience caused by tickling themselves just as intense as that occurring when they are tickled by someone else (Blakemore *et al.* 2000, and Lindner *et al.* 2005 in relation to eye movements, see also Shergill *et al.* 2005 in relation to the sense of pressure). We believe that these observations give us clue about what it feels like to have the experience labelled *delusion of control* (Hohwy & Frith 2004). Because of the failure to attenuate sensory feedback associated with the movement, voluntary movements actually feel like involuntary movements. In other words, it doesn't feel like a movement that has been caused by my intention to move.

4. The experience of agency: expectations and outcomes

This account of the delusion of control assumes that an abnormal experience (*failure to attenuate sensory feedback during a voluntary movement*) is

sufficient to create an abnormal belief (*believing that external forces are causing the movements*). But several studies suggest that this is not the case. Patients with delusions do not simply have an abnormal sensory experience; they also have an abnormal experience of agency, the experience that I am the cause of this movement and its consequences. Intensive study of this experience by Patrick Haggard and colleagues has revealed the phenomenon of *intentional binding* (Haggard *et al.* 2002). When we intend to perform an action, that action and its consequences are experienced as being closer together in mental time than they are in physical time. This binding together of actions and their intended consequences has both a predictive and a postdictive component (Haggard & Chambon 2012). The binding effect is greater when the outcome of the action is more strongly expected, occurring 75% of the time rather than 50% of the time. This is the predictive component. On the other hand, the perceived time of making a movement is altered if the expected consequence of making that movement does not subsequently occur. This is a postdictive effect.

Two recent studies of patients with delusions of control have shown that the experience of agency in these patients depend largely on the outcome of the movement (postdictive effect) and very little upon expectations (predictive effect). In the first experiment (Voss *et al.* 2010) these effects were measured directly using the methods developed by Haggard. From this paradigm it appeared that patients showed no predictive component for their awareness of action and an abnormally large retrospective component. In other words their experience of action was almost entirely determined by the outcome of the action. The second experiment (Synofzik *et al.* 2010) used a very different technique in which subjects had to make pointing movements in a virtual reality setup where they were given distorted visual feedback about the position of their hand. For example, if they pointed straight ahead they would consistently see their pointing movement rotated five degrees to the left. With this paradigm it is also possible to distinguish between the role of expectations and outcomes. From their ability to detect the visual feedback rotations it was shown that the patients' motor expectations were less precise than those of the controls. At the same time the patients' pointing behaviour adapted to the false visual feedback better than the controls demonstrating a greater reliance on movement outcomes. The size of both these effects correlated with the severity of the patients' delusions of control. These results suggest that patients with delusions of control have problems combining information from two different sources: that is from prior expectations about motor movements and subsequent outcomes of motor movements. They put too much weight on outcome and too little

on expectations. This result gives us a deeper understanding of why patients are abnormally aware of the consequences of their actions.

5. A Bayesian approach

In the normally functioning brain information from different sources is combined in a statistically optimum manner (e.g. Ernst & Banks 2002). The mechanism for achieving this is well captured in a Bayesian framework (Kersten *et al.* 2004, Yuille & Kersten 2006). When we perform an action we predict the immediate outcome of the action on the basis of our prior knowledge. If the outcome is not what we expect (a prediction error) then we modify the knowledge on which our expectation was based and this updated knowledge determines our future expectations. All this happens at a sub-personal level. That is to say, we are not consciously aware of prior expectations, prediction errors, or updating except, perhaps, when the prediction error is large. In terms of neuronal representations, precision can be thought of as amplifying prediction errors associated with a high degree of certainty or reliability. Crucially, we also need to update predictions about the precision of prediction errors. These expectations encode our uncertainty or confidence about predictions – irrespective of their content (the expected precision is sometimes referred to as expected [un]certainty).

More generally, Bayes' theorem (Bayes 1763/1958) provides a measure of the extent to which some new evidence (e.g. the prediction error) requires that we update our beliefs about the world. Within this framework there is no qualitative distinction between perception and belief, since both involve making inferences about the state of the world on the basis of evidence (Fletcher & Frith 2009). In the case of perception, this is the evidence of our senses (Helmholtz 1878). The framework also indicates the statistically optimal procedure for combining evidence from different sources. The different sources of evidence should be weighted by their precision (the inverse of variability), with the more precise evidence being given the greater weight. Likewise, if our belief (prior knowledge) about the world is assigned a greater precision, a much greater quality of evidence will be needed before we up-date it.

Nevertheless, evidence from a very precise source, such as vision, can alter what might be expected to be well-established beliefs that are held with high precision. An example of this is the *rubber hand* illusion (Botvinick & Cohen 1998). To create this illusion the participant sits at a table with one arm out of sight under a shelf. On top of the shelf is placed a prosthetic limb roughly lined up with the real arm. The experimenter then synchronously strokes the real hand and the rubber hand. Within about one minute,

participants have the vivid experience that the rubber hand is now their own hand. Objective evidence for this experience can obtained by threatening the rubber hand which elicits a physiological responses (e.g. Ehrsson *et al.* 2004) and from asking the participant to make aiming movements which indicate that the participant is representing the position of the rubber hand as the starting point for a movement rather than that of the real hand (e.g. Chambon *et al.* 2012).

A Bayesian interpretation of this effect is as follows. Before the development of the illusion, a participant experiences highly synchronised stimuli in vision and touch which seem to come from different spatial locations, the rubber hand and the real hand respectively. To resolve this discrepancy the location of the touch stimulation is "moved" to coincide with the location of the visual stimuli on the rubber hand. In this example, the prior expectation that synchronised stimuli come from the same location and the precision of the visual sense with regard to spatial location over-ride the evidence from the somewhat less precise tactile and kinaesthetic senses.

Using this framework we can model some different ways in which false perceptions and false beliefs might arise (Corlett *et al.* 2009). For example, if too much weight was put on the evidence, i.e. the prediction errors, then people would be constantly up-dating their beliefs about the world, but never fully resolving the problem. At the other extreme, if too much weight was put on prior expectation, then people would see only what they expected to see. In extreme cases, this would lead to perception without any sensory input, resulting in hallucinations. In principle, the different models that can arise in the Bayesian framework might relate to the different forms of hallucinations and delusions associated with different disorders and different pharmacological treatments. Furthermore, the different kinds of illusion to which we are all subject will have different causes in terms of the model.

The rubber hand illusion arises because the discrepancy in the location of tactile and visual sensations is treated as a prediction error, which is eliminated by assuming that the felt real hand is at the same location as the seen rubber hand. Patients with schizophrenia acquire this illusion more rapidly and strongly than control participants (Peled *et al.* 2000), presumably because they put even more weight on the apparent prediction error. The *hollow face* illusion, in contrast, arises because too much weight is put on prior expectations. From our extensive experience with faces, we know that the nose sticks out in front. But if we look at a hollow mask of a face from the back (i.e. a concave face), then this expectation is not fulfilled since the nose is the part of the face that is furthest from us. In this case our expectations override the evidence of our sense and we see a normal convex face. Here

patients with schizophrenia are *less* susceptible to the illusion than control participants (Koethe *et al.* 2006). Thus, in both these examples, as with their perception of agency, patients with schizophrenia put more weight on sensory evidence (prediction errors) and less weight on prior expectations.

6. The role of dopamine in the generation of perceptions and beliefs

As we mentioned at the beginning of this essay, it has long been established that the neurotransmitter dopamine is implicated in the generation of hallucinations and delusions. But it is only recently that we are beginning to understand the precise nature of this role (Corlett *et al.* 2009, Kapur 2003, Stephan *et al.* 2009). The breakthrough came with the demonstration by Wolfram Schultz and colleagues that activity in dopamine-containing neurons could be seen as a signal of reward prediction error (Schultz & Dickinson 2000), where later work highlighted the role of dopamine in reporting the certainty or predictability of a reward; namely, the precision of reward prediction errors (Schultz *et al.* 2008).

Prior to this observation, activity in these neurons was seen as a signal of reward, since activity increased immediately after an animal received a reward, for example a drink of juice. Schultz and colleagues used Pavlovian conditioning paradigms in which animals learned that the reward would arrive one second after a visual cue. Before learning had occurred increased neural activity occurred immediately after presentation of the juice. However, after learning had occurred there was no response to the presentation of the juice, but there was a response to the presentation of the cue. These observations fit with the idea that the activity occurs when there is an unexpected signal of reward, i.e. a positive prediction error. When the reward arrives at the expected time after the cue, then there is no prediction error and no activity. In contrast, the animal does not know when the cue is going to arrive. So the cue now creates a positive prediction error. If, after learning, the reward was omitted, there was a reduction of neural activity, consistent with a negative prediction error, since the expected reward did not arrive.

Prediction errors can be used to continuously up-date representations of an ever-changing world. This process can be studied in simple probabilistic learning tasks. For example, the participant has to learn that choice A will be rewarded 80% of the time, while choice B is rewarded 20% of the time. Before learning starts, the two options will have roughly equal value. When a choice is rewarded, this creates a positive prediction error and the value of that option is increased. When a choice is not rewarded the value of that option is decreased. After some experience the participants' internal representations of the value of the options will reflect the reward

probability of these options (Sutton & Barto 1998). The rate at which the subjects learns depends upon the precision of reward prediction errors (Mathys *et al.* 2011) and should therefore depend upon manipulations of expected precision in the brain:

The rate of this kind of learning can be modified by manipulating the dopamine system. For example, Mathias Pessiglione and colleagues (2006) treated human volunteers with L-DOPA or haloperidol, drugs which respectively activate or deactivate the dopamine system, while the participants performed a simple probabilistic learning task. Activation of the dopamine system caused faster learning, while deactivation caused slower learning, although, interestingly, the effect only applied to learning about gains, not losses. This study, along with others, specifies a role for dopamine in probabilistic learning.

There is much evidence that this kind of learning, in which representations about the state of the world (beliefs) are up-dated on the basis of new evidence, is disrupted in schizophrenia. For example, many studies (e.g. Garety *et al.* 1991) have found that patients with schizophrenia "jump to conclusions", in that they base their conclusions on less evidence than controls. There is also evidence for abnormalities in the integration of new evidence into beliefs (Freeman *et al.* 2002) and for a bias against disconfirmatory evidence (Woodward *et al.* 2008).

Traditionally, the delusions, or false beliefs, associated with schizophrenia have been assumed to reflect a defect in reasoning. However, as anyone who has argued with patients about their delusions will have experienced, their logic can be impeccable. The studies listed above suggest that the reasoning problem associated with delusions may be restricted to probabilistic reasoning (see for example Howes *et al.* 2007). When logical reasoning is investigated patients show little abnormality (Kemp *et al.* 1997, Owen *et al.* 2007) or may even perform better than controls (Mellet *et al.* 2006).

At the physiological level there is also evidence for abnormalities in schizophrenia relating more specifically to prediction errors. When performing tasks that elicited *reward* prediction errors (Murray *et al.* 2007) or *causal inference* prediction errors (Corlett *et al.* 2007) schizophrenic patients were observed to show less response to such errors in the dopamine rich areas of the mid-brain.

7. How false prediction errors generate false beliefs

Given the evidence that schizophrenia is associated with abnormal probabilistic learning, linked to abnormal modulation of prediction errors and an over-active dopamine system, We shall now speculate on how different

kinds of failure in the prediction error system might lead to false perceptions and false beliefs. As we have seen, delusions of control, in which patients believe that their actions are caused by some external force, are associated with a failure to suppress the sensory consequences of a self-generated movement. This is an example of a falsely attenuated prediction error

What is it that can go wrong, in neural terms, with the Bayesian mechanism that could create false predictions? First, there is the possibility that, through loss of neural connections, prediction error signals are not generated (or selectively enabled by a high precision) when there actually is an error. As a result beliefs are not updated when they should be. This seems to be the case for neurological patients with *anosognosia* (Schultz *et al.* 2008). This disorder is typically associated with damage to the right parietal cortex associated with stroke. Such patients falsely believe that they can and do move their paralysed left limb. There is evidence that this disorder can be explained as follows (Fotopoulou *et al.* 2008). In the normal case, as outlined in section 3 of this essay, the intention to move creates a prediction of the consequences of the movement, both in terms of the future position of the limb and the sensory consequences of the movement. It is predictions that dominate our awareness of the action we are making. These predictions are compared with the actual outcome of the action. If there are discrepancies, prediction errors are generated which alter the representation of the action. In the case of anosognosia, the motor system concerned with the intention to move is intact and predictions are generated. But, due to the right parietal damage, there is no signal concerning the outcome of the action and no prediction error is generated. In consequence patients continue to believe that they have moved their limb.

A second possibility is that, through loss of neural connections, a precise prediction error is generated when it is inappropriate. This seems to be the case for patients with Capgras syndrome (Capgras & Reboul-Lachaux 1923). These patients falsely believe that a familiar person, typically the spouse, has been replaced by a double. The assumption here is that face recognition has a cognitive and an emotional component. Via the cognitive component we discover the identity of the face we are looking at. At the same time and independently, an emotional response is generated if the face belongs to a familiar person. Thus, when a face is identified as familiar an emotional response is expected. In the case of Capgras syndrome, probably through damage to the amygdala or its connections, there is no emotional response to a familiar face. This discrepancy between the identity signal and the emotion signal creates a prediction error. As a result the patient's belief about the identity of the face is inappropriately updated. "*This person looks*

like my wife, but there is something not quite right about her. It can't be my wife, but someone who looks like her" (Ellis & Young 1990).

In each of these two examples a circumscribed false belief was associated with circumscribed brain damage. In the case of schizophrenia, the false beliefs typically involve many different domains and often become more widespread with time. If these delusions also result from false predictions, then the abnormality is not likely to result from circumscribed neural disconnections. This expectation is consistent with the evidence that the dopamine system is involved. Abnormalities of this system have an impact on many brain regions.

How might predictions become false in the absence of the kinds of disconnections discussed above? Within a Bayesian framework (see section 5) precision is a very important property of a signal. More weight is given to signals with high precision. Karl Friston and his colleagues have proposed that dopamine controls the precision of prediction errors (Friston *et al.* 2012). If, as a result of excessive dopaminergic activity, prediction errors became abnormally precise, then beliefs would become updated on the basis of signals that would normally have been ignored (see Kapur 2003 for a closely related version of this idea).

A general effect whereby prediction errors became abnormally precise would have an impact on many domains. In addition, the long-term experience of false prediction errors might cause patients to put less and less weight on their prior expectations. This is because these expectations would persistently be signalled as being wrong. This formulation fits nicely with the observations discussed in section 4, showing that the experience of agency in patients with delusions of control depends upon less weight being given to expectations and more weight being given to the outcomes of motor movements.

8. A hierarchy of beliefs

There is an obvious problem with this account of the generation of false beliefs in schizophrenia. If they put too much weight on new evidence and too little weight on prior beliefs, then we would expect that they should be constantly changing their beliefs. This is clearly not the case. While the scope of their false beliefs may be slowly modified over time, the striking feature of delusions is that patients will stick with them despite what is perceived by everyone else as very good evidence against their belief. We suggest this problem can be resolved if we recognise that perceptions and beliefs do not exist in isolation, but are developed within a hierarchy. It is the beliefs at the top of this hierarchy that are particularly resistant to change.

In Karl Friston's account of these Helmholtzian ideas (Friston 2005), the brain uses a hierarchy of predictions, where expectations at any level provide prior beliefs for the level below (these are known as empirical priors in statistics). Each level integrates new evidence from the level below and (empirical) prior expectations from the level above to generate a prediction error. This prediction error is fed upwards as the *evidence* for the next level of the hierarchy. Likewise, the prior expectations at the higher levels of the hierarchy (empirical priors) are fed downwards to constrain the possible explanations of the prediction errors coming from the lower levels. Crucially, the weights assigned to bottom-up prediction errors and top-down predictions depend upon the relative precisions (possibly encoded by dopamine) at each level of the hierarchy.

The lowest level of this hierarchy of perceptions and beliefs is the most closely linked to raw sensation, while the higher levels are concerned with more abstract levels of representation. The process of reading provides a useful illustration of the workings of such a hierarchy. At a low level we have the graphic shape components of which the letters are composed, and then we can move up the hierarchy through representations of words and sentences, reaching meaning at the highest level. However, reading is not a linear process, moving steadily upwards from shapes to meanings. The high level of meaning will constrain how signals are interpreted at the low level of shape. Consider, for example, the string of shapes event. The ev in this string is ambiguous and could be seen as "w" or as "ev". How it is seen will depend on the meaning of the sentence in which it occurs: "w" in "Jack and Jill *went* up the hill", "ev" in "the pole vault was the last *event*". The meaning of the sentence has had a top-down effect on our perception at a much lower level of the hierarchy. Presented with these two sentences we will read them easily without ever noticing the ambiguity of the shapes used to write them.

If the prediction errors being generated at the bottom of this hierarchy are treated as being unduly precise, then their effects will gradually work upwards through the hierarchy, and they will never be fully eliminated by changing low-level beliefs about the world. Consider what might happen if something goes wrong with the fancy system in my car that signals problems. In particular, assume that an error warning light is unduly sensitive to fluctuations in the engine's performance from normal levels. This would correspond to a pathologically highly precision at the sensory level, leading to a dashboard warning light that is almost continuously illuminated. I am led to falsely believe that there is indeed something wrong with the engine. I take my car to the garage and they report that nothing is wrong. However,

the light is still on and keeps on signalling an error. So, this leads me to falsely believe that the garage is incompetent. I report them to the "good garage guide" who investigate and conclude that the garage is not incompetent. Now I believe that the "good garage guide" is corrupt.

We suggest that, in the case of schizophrenia, it is the beliefs at the top of the hierarchy that are so resistant to change. This is because, for the patient, they seem to be the only way of explaining away the apparent problems with lower levels of the hierarchy. In the normal case prediction errors at the lowest level of the hierarchy elicit changes in our interpretation of sensory input. This enables us to develop an increasingly accurate account of the causes of our sensations. In other words, we develop a representation of the world that corresponds ever more closely to reality. Falsely precise prediction errors undermine this process and lead us ever further from reality. They are transmitted up our hierarchy of beliefs as we attempt to explain them away (Fletcher & Frith 2009).

This process has been described in a particular striking manner by Peter Chadwick (1993). Chadwick, who has a PhD in psychology, has described in detail his experiences during an episode of paranoid schizophrenia. In my opinion, his descriptions lend themselves well to the account of delusions we have developed in this essay. He says, "*I had to make sense, any sense, out of all these uncanny coincidences. I did it by radically changing my conception of reality*". In our terminology, these uncanny coincidences were false hypotheses engendered by prediction errors with inappropriately high precision or salience. To explain them away Chadwick had to conclude that other people, including radio and television presenters, could see into his mind. This was the radical change he had to make in his conception of reality.

9. Conclusions

We suggest that the Bayesian framework, outlined here, for explaining perceptions and beliefs provides a plausible account of the development of hallucinations and delusions in schizophrenia. In addition, the account can be directly linked to physical processes involving the dopamine system of the brain. In principle, such an account can provide a guide for the development of new treatments, whether these are at the cognitive or the biological level. For example, it might be possible to develop a method for reducing the precision of prediction errors.

Explorations of abnormal behaviour and experience will always illuminate our general understanding of the mind. This account of the generation of false beliefs in the case of schizophrenia, makes me realise how fragile this process is and how easily it might go astray in the normal case. Given

the right anomalous sensory experiences each of us could develop some bizarre and erroneous belief system. Why does this not happen more often? We believe that we are usually saved from taking such erroneous paths by the constraints provided by those even higher levels of the belief hierarchy that are external to our brains. These constraints arise from our interactions with our peers and with our culture. Even at the lowest perceptual level of the hierarchy the high level constraints that arise from interactions with others enable us to achieve accounts of the world that are more accurate than those that we can develop on our own (Bahrami *et al.* 2010). It is this submission of our own ideas to the criticism of others that has been formalised in the practice of science.

In the case of schizophrenia, in contrast, these high level external constraints no longer seem to operate. Patients stick to their false beliefs in spite of the objections of others. Is this an inevitable consequence of the process by which prediction errors filter up through the hierarchy, or is this evidence for some additional problem that needs to be identified? Further research is needed.

Acknowledgements

Our work is supported by the Wellcome Trust, the Danish National Research Foundation and Aarhus University. We are grateful to Uta Frith and Rosalind Ridley for their comments.

References

Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD. 2010. Optimally interacting minds. *Science* 329: 1081-5.

Bayes T. 1763/1958. Studies in the History of Probability and Statistics: IX. Thomas Bayes' Essay Towards Solving a Problem in the Doctrine of Chances. *Biometrika* 45: 296-315.

Blakemore SJ, Frith CD, Wolpert DM. 1999. Spatio-temporal prediction modulates the perception of self- produced stimuli. *J. Cogn. Neurosci.* 11: 551-59.

Blakemore SJ, Smith J, Steel R, Johnstone EC, Frith CD. 2000. The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: evidence for a breakdown in self-monitoring. *Psychol. Med.* 30: 1131-39.

Botvinick M, Cohen J. 1998. Rubber hands 'feel' touch that eyes see. *Nature* 391: 756.

Capgras J, Reboul-Lachaux J. 1923. L'illusion de 'sosies' dans un délire systématise chronique. *Bull. Soc. Clin. Med. Ment.* 2: 6-16

Chadwick PK. 1993. The step ladder to the impossible: A first hand phenomenological account of a schizo-affective psychotic crisis. *Journal of Mental Health* 2: 239-50.

Chambon V, Wenke D, Fleming SM, Prinz W, Haggard P. 2012. An Online Neural Substrate for Sense of Agency. *Cereb Cortex*.

Connell P. 1958. *Amphetamine Psychosis*. London: Chapman & Hall.

Corlett PR, Frith CD, Fletcher PC. 2009. From drugs to deprivation: a Bayesian framework for understanding models of psychosis. *Psychopharmacology (Berl)* 206: 515-30.

Corlett PR, Murray GK, Honey GD, Aitken MR, Shanks DR, *et al*. 2007. Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain* 130: 2387-400.

Ehrsson HH, Spence C, Passingham RE. 2004. That's My Hand! Activity in Premotor Cortex Reflects Feeling of Ownership of a Limb. *Science*: 1097011.

Ellis HD, Young AW. 1990. Accounting for delusional misidentifications. *British Journal of Psychiatry* 157: 239-48.

Ernst MO, Banks MS. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415: 429-33.

Fletcher PC, Frith CD. 2009. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat Rev Neurosci* 10: 48-58.

Fotopoulou A, Tsakiris M, Haggard P, Vagopoulou A, Rudd A, Kopelman M. 2008. The role of motor intention in motor awareness: an experimental study on anosognosia for hemiplegia. *Brain* 131: 3432-42.

Freeman D, Garety PA, Kuipers E, Fowler D, Bebbington PE. 2002. A cognitive model of persecutory delusions. *Br J Clin Psychol* 41: 331-47.

Friston K. 2005. A theory of cortical responses. *Philos T Roy Soc B* 360: 815-36.

Friston KJ, Shiner T, FitzGerald T, Galea JM, Adams R, *et al*. 2012. Dopamine, affordance and active inference. *PLoS computational biology* 8: e1002327.

Frith C. 2002. Attention to action and awareness of other minds. *Conscious Cogn* 11: 481-7.

Garety PA, Hemsley DR, Wessely S. 1991. Reasoning in deluded schizophrenic and paranoid patients. Biases in performance on a probabilistic inference task. *J. Nerv. Ment. Dis.* 179: 194-201.

Haggard P, Chambon V. 2012. Sense of agency. *Curr Biol* 22: R390-2

Haggard P, Clark S, Kalogeras J. 2002. Voluntary action and conscious awareness. *Nat. Neurosci.* 5: 382-85.

Helmholtz Hv. 1866. *Handbuch der Physiologischen Optik*. Leipzig: Voss.

Helmholtz Hv. 1878. The Facts of Perception In *Selected Writings of Hermann Helmholtz*: Wesleyan University Press.

Hohwy J, Frith C. 2004. Can neuroscience explain consciousness? *J. Conscious. Stud.* 11: 180-98.

Howes OD, Montgomery AJ, Asselin MC, Murray RM, Grasby PM, McGuire PK. 2007. Molecular imaging studies of the striatal dopaminergic system in psychosis and predictions for the prodromal phase of psychosis. *Br J Psychiatry Suppl* 51: s13-8.

Johnstone EC, Crow TJ, Frith CD, Carney MW, Price JS. 1978. Mechanism of the antipsychotic effect in the treatment of acute schizophrenia. *Lancet* 1: 848-51.

Kambeitz-Ilankovic L, Hennig-Fast K, Benetti S, Kambeitz J, Pettersson-Yeo W, *et al*. 2012. Attentional Modulation of Source Attribution in First-Episode Psychosis: A Functional Magnetic Resonance Imaging Study. *Schizophr Bull*.

Kapur S. 2003. Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry* 160: 13-23.

Kemp R, Chua S, McKenna P, David A. 1997. Reasoning and delusions. *Br J Psychiatry* 170: 398-405.

Kersten D, Mamassian P, Yuille A. 2004. Object perception as Bayesian inference. *Annu Rev Psychol* 55: 271-304.

Koethe D, Gerth CW, Neatby MA, Haensel A, Thies M, *et al*. 2006. Disturbances of visual information processing in early states of psychosis and experimental delta-

9-tetrahydrocannabinol altered states of consciousness. *Schizophr Res* 88: 142-50.

Lindner A, Thier P, Kircher TT, Haarmeier T, Leube DT. 2005. Disorders of agency in schizophrenia correlate with an inability to compensate for the sensory consequences of actions. *Curr Biol* 15: 1119-24.

Mathys C, Daunizeau J, Friston KJ, Stephan KE. 2011. A bayesian foundation for individual learning under uncertainty. *Front Hum Neurosci* 5: 39

Mellet E, Houde O, Brazo P, Mazoyer B, Tzourio-Mazoyer N, Dollfus S. 2006. When a schizophrenic deficit becomes a reasoning advantage. *Schizophr Res* 84: 359-64.

Mellor CS. 1970. First rank symptoms of schizophrenia. *British Journal of Psychiatry* 117: 15-23.

Murray GK, Corlett PR, Clark L, Pessiglione M, Blackwell AD, *et al*. 2007. Substantia nigra/ventral tegmental reward prediction error disruption in psychosis. *Molecular Psychiatry* 13(3): 267-76.

Owen GS, Cutting J, David AS. 2007. Are people with schizophrenia more logical than healthy volunteers? *Br J Psychiatry* 191: 453-4.

Peled A, Ritsner M, Hirschmann S, Geva AB, Modai I. 2000. Touch feel illusion in schizophrenic patients. *Biological Psychiatry* 48: 1105-8.

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442: 1042-5.

Schultz W, Dickinson A. 2000. Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* 23: 473-500.

Schultz W, Preuschoff K, Camerer C, Hsu M, Fiorillo CD, *et al*. 2008. Explicit neural signals reflecting reward uncertainty. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 363: 3801-11.

Shergill SS, Samson G, Bays PM, Frith CD, Wolpert DM. 2005. Evidence for sensory prediction deficits in schizophrenia. *American Journal of Psychiatry* 162: 2384-6.

Stephan KE, Friston KJ, Frith CD. 2009. Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. *Schizophr Bull* 35: 509-27.

Sutton RS, Barto AG. 1998. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Synofzik M, Thier P, Leube DT, Schlotterbeck P, Lindner A. 2010. Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions. *Brain* 133: 262-71.

Voss M, Moore J, Hauser M, Gallinat J, Heinz A, Haggard P. 2010. Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences. *Brain* 133: 3104-12.

Woodward TS, Moritz S, Menon M, Klinge R. 2008. Belief inflexibility in schizophrenia. *Cognit Neuropsychiatry* 13: 267-77.

Yuille A, Kersten D. 2006. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn Sci* 10: 301-8.

Addiction: A Disease of Self-Control

Nora D. Volkow and Ruben Baler

Introduction

Research on the neuroscience of substance use disorders (SUDs) has started to shed light on the ways in which chronic drug abuse changes the brain to cause the profound disruptions we see in the behavior of an addicted person. This is because drugs of abuse impact many neuronal circuits that are crucial for the orchestration of *conscious experience* and hence, for the proper (flexible) functioning in social environments. But the core impairments established during the addictive process have a particularly devastating impact on the interacting circuits of motivational drive (which is enhanced for drug-related stimuli) and of self-control (which is weakened by chronic drug exposure) (Kalivas and Volkow 2005).

For many years, studies of addiction had focused mainly on the role of dopamine and the brain reward circuitry (Di Chiara 1999; Weiss and Koob 2001). However, the drug-induced DA boost fails to fully explain addiction since it happens in naïve animals and its magnitude is decreased in addiction (Volkow *et al.* 1997). Preclinical and clinical studies are revealing neuroadaptations in frontocortical regions of the brain that are likely to underlie compulsive drug-seeking behaviors in addiction (for review, see (Goldstein and Volkow 2002)). Imaging studies have provided particularly compelling evidence for the involvement of the brain's cognitive system (i.e., prefrontal cortex [PFC], including orbitofrontal cortex [OFC], anterior cingulate gyrus [ACC]) in the addiction process (Volkow and Fowler 2000). More recent work has revealed that the PFC plays a crucial role in social cognition and emotions (Forbes and Grafman 2010), which are key to proper social integration including responses to social rewards and punishment. For example, damage to ventral areas, frontal, and striatal regions can interfere with the ability of a person to accurately distinguish right from wrong in a socially acceptable manner, which can lead to socially inappropriate behaviors (Koenigs *et al.* 2007). Similarly, such brain impairments decrease the sensitivity to social rewards or punishments that can lead a person to behave in ways that alienate others and result in social isolation or even incarceration. While repeated drug exposures impair the fronto-striatal circuit there is also evidence that genetics or social adverse environmental exposures during childhood/adolescence can also result in impairments in this circuit that increase the vulnerability of the individual for a substance use disorder.

Drugs' impact on the neural substrate of cognition

Humans addicted to drugs display a significant reduction in dopamine receptor type 2 (D2R) function in the striatum (including the Nucleus Accumbens [NAc] located in the ventral striatum), an effect that has been implicated in impulsive and compulsive behavioral phenotypes (Volkow *et al.* 2012). In the human brain, the reductions in D2R in striatum are associated with decreased activity in the OFC (including the right inferior cortex, which is necessary for inhibition), anterior cingulate cortex (ACC), and dorsolateral prefrontal cortex (DLPFC) (Volkow *et al.* 2001;Volkow *et al.* 1993; Volkow *et al.* 2007) indicating the impulsive/compulsive phenotypes reflect the impaired modulation by the D2R striato-cortical pathway, which is inhibitory, in PFC. Studies have also shown decreased frontal cortical activity during intoxication from many drugs of abuse (Chang and Chronicle 2007) and disruption of several frontocortical processes with chronic drug use (Table I) (see (Goldstein and Volkow 2012) for a review). Predictably, targeting the frontal impairments in addiction has been proposed as a therapeutic strategy to improve self-control (Goldstein *et al.* 2010;Volkow *et al.* 2013).

Among the frontal regions implicated in addiction, the OFC, ACC, and DLPFC participate in salience attribution, inhibitory control/emotion regulation and error detection, and decision making, respectively. It has been postulated that their improper regulation by striatal D2R-mediated DA signaling in addicted subjects could underlie the enhanced motivational value of drugs and the loss of control over drug intake (Volkow and Fowler 2000). Incidentally, related dysfunctions could also underlie behavioral addictions, like pathological internet use (Yuan *et al.* 2012) and compulsive overeating in some cases of obesity (Volkow *et al.* 2012). In parallel, investigators have also uncovered differential modulation of reward-seeking behavior by D1R versus D2R in the PFC. For example, recent preclinical studies have shown that pharmacologic blockade of mPFC D1R attenuates; whereas D2R increases a tendency for risky choices, providing evidence for a dissociable but complementary role of medial PFC DA receptors that is likely to play a major role in orchestrating the fine balance needed for inhibitory control, delayed discounting, and judgment (St Onge *et al.* 2011).

In addition, because impairments in OFC and ACC are associated with compulsive behaviors and impulsivity, DA's impaired modulation of these regions is likely to contribute to the compulsive and impulsive drug intake seen in addiction (Volkow and Fowler 2000). Clearly, low DA tone in PFC could also represent a preexisting vulnerability for drug use, albeit one that is likely to be exacerbated with the further decreases in striatal D2R associated with drug addiction. Indeed, a study performed in subjects who, de-

spite a positive family history (high risk) of alcoholism, were not themselves alcoholics, revealed a higher than normal striatal D2R availability that was associated with normal metabolism in OFC, ACC, and DLPFC (Volkow *et al.* 2006). This suggests that, in these subjects at risk for alcoholism, the normal prefrontal function was linked to enhanced striatal D2R signaling, which in turn may have protected them from alcohol abuse.

The central role of the PFC among the neural targets of addictive drugs may also help explain why addiction is a developmental disease whose chances of becoming expressed are increased if drug use onset takes place during childhood or adolescence. The heavy bidirectional connectivity between the PFC and limbic regions is instrumental in directing affective and social behaviors and is not fully developed until young adulthood. The maturation of fronto-limbic connectivity is highly sensitive to the deleterious impact of environmental factors such as chronic stress, parental neglect, drugs, and social experiences (Kolb *et al.* 2012). This makes the PFC susceptible to abnormal developmental trajectories, which can increase the risk for addiction and other psychiatric disorders.

Cognitive and behavioral implications

Through a fascinating but sinister process, drugs disrupt the very neurobiological systems underpinning the assessment of what's *important* in a person's life. From a biological perspective, we think much of the addictive behavior phenotype can be explained by the ability of chronic drug exposure to cause neuroadaptations in brain reward and control systems, including the emergence of conditioned associations that link the rewarding experience from the drug to the multiple cues that surround it. In this way, drug addicted individuals suffer from a profoundly distorted system of value placement, which can devastate their self-determination capacity. The structural and functional changes that accompany these drug-induced dysfunctions are long lasting, and can persist even after years of drug discontinuation, which is one of the main reasons why we define addiction as a chronic and relapsing disease of the brain.

Furthermore, while the value that an addicted individual places on drug reward becomes unsustainably exaggerated, the potential impact of deleterious consequences (e.g., familial dislocation, becoming the target of drug-related violence, or incarceration) becomes progressively devalued. The establishment of such a severe imbalance in how an addicted individual attributes value to both rewarding and aversive situations and stimuli has a profound and negative impact on the individual's social competence. His/her behaviors are now governed by the uncontrollable overvaluing of

the drug (enhanced expectation of a positive reward) and by a growing insensitivity to the deterrent value of potential punishments (reduced fear of a negative reward). The problem is further compounded by the tendency of many substance abusers, more so than nonusers, to routinely choose immediate rewards over delayed gratification (e.g., choose $20 dollars now rather than wait 1 week in order to get double that amount) an impairment associated with dysfunction of ventral prefrontal regions (Ernst and Paulus 2005). This inability to appropriately weigh delayed rewards can be devastating to an addicted person who may be willing to sacrifice future gains or incur major losses in exchange for instant gratification. An individual in this situation may not think twice about the risk of losing his or her freedom tomorrow in order to chase the high from the drug today. This knowledge helps explain why the prevailing social system that dangles some future threat of imprisonment over an addict's head seldom deters immediate substance abuse-related behaviors in addicted subjects. It also highlights the need to provide addicted individuals with alternative reinforcers as a strategy both for the prevention of SUD as well as for its treatment.

Implications for treatment and social policy

Behavioral inhibition is fundamental to the success of social intercourse, which is critically dependent on a person's ability to control impulsive behaviors whenever this is needed. For most people, the combination of biological (e.g., individual-level characteristics) and environmental (e.g., culture, laws, religion) factors build up a sufficiently robust mechanism to inhibit or at least help manage internally or externally generated temptations. But the system is not fail-safe and some individuals at one extreme of the impulsivity spectrum, as is the case in addiction, are the constant victims of powerful, unstoppable urges. By perturbing the function of the PFC, the addiction process degrades the very substrates that enable an individual to make appropriate decisions, exercise self-determination and exert free will. This is actionable knowledge that can and should be parlayed into more effective treatments. Therapeutic interventions should create incentives for the substance abusers to engage and stay in treatment; including strategies that help strengthen social ties with family and community, for social interactions are powerful reinforcers that can provide the addicted individual with alternatives that could help counteract the high-reward value of drugs.

An important consequence of the long-term brain adaptations discussed above is that most addicted patients will require a long period of treatment, during which relapse is likely to occur, thus relapse must be considered a predictable setback and not a failure of the treatment. This also explains

why the best treatment outcomes are reported by programs that offer continuity of care for a 5-year period (McLellan *et al.* 2008). It is equally important to recognize that social isolation is a well-recognized risk factor of mental illness including addiction (Karelina and DeVries 2011; Seo and Huang, 2012; Simoni-Wastila and Yang 2006). Yet, by most accounts, stigmatization and/or incarceration have been society's prevailing responses to addicted individuals. Such stigmatization impedes the search for treatment and further isolates addicted individuals and their families.

Finally, the implications derived from the current understanding of addiction could be easily misconstrued as advocating a sort of moral relativism at the expense of individual responsibility. Yet nothing could be farther from the truth; for the addicted individual is responsible for the management and

Process	Possible disruption in addiction	Probable PFC region
Self-control and behavioural monitoring: response inhibition, behavioural coordination, conflict and error prediction, detection and resolution	Impulsivity, compulsivity, risk taking and impaired self-monitoring (habitual, automatic, stimulus-driven and inflexible behavioural patterns)	DLPFC, dACC, IFG and vlPFC
Emotion regulation: cognitive and affective suppression of emotion	Enhanced stress reactivity and inability to suppress emotional intensity (for example, anxiety and negative affect)	mOFC, vmPFC and subgenual ACC
Motivation: drive, initiative, persistence and effort towards the pursuit of goals	Enhanced motivation to procure drugs but decreased motivation for other goals, and compromised purposefulness and effort	OFC, ACC, vmPFC and DLPFC
Awareness and interoception: feeling one's own bodily and subjective state, insight	Reduced satiety, 'denial' of illness or need for treatment, and externally oriented thinking	rACC and dACC, mPFC, OFC and vlPFC
Attention and flexibility: set formation and maintenance versus set-shifting, and task switching	Attention bias towards drug-related stimuli and away from other stimuli and reinforcers, and inflexibility in goals to procure the drug	DLPFC, ACC, IFG and vlPFC
Working memory: short-term memory enabling the construction of representations and guidance of action	Formation of memory that is biased towards drug-related stimuli and away from alternatives	DLPFC
Learning and memory: stimulus–response associative learning, reversal learning, extinction, reward devaluation, latent inhibition (suppression of information) and long-term memory	Drug conditioning and disrupted ability to update the reward value of non-drug reinforcers	DLPFC, OFC and ACC
Decision making: valuation (coding reinforcers) versus choice, expected outcome, probability estimation, planning and goal formation	Drug-related anticipation, choice of immediate reward over delayed gratification, discounting of future consequences, and inaccurate predictions or action planning	lOFC, mOFC, vmPFC and DLPFC
Salience attribution: affective value appraisal, incentive salience and subjective utility (alternative outcomes)	Drugs and drug cues have a sensitized value, non-drug reinforcers are devalued and gradients are not perceived, and negative prediction error (actual experience worse than expected)	mOFC and vmPFC

Table 1. Processes associated with the prefrontal cortex that are disrupted in addiction. Orbitofrontal cortex (OFC) includes Brodmann area (BA) 10-14 and 47, and inferior and subgenual regions of anterior cingulate cortex (ACC) (BA 24, 25 and 32) in the ventromedial prefrontal cortex (vmPFC); ACC includes rostral ACC (rACC) and dorsal ACC (dACC) (BA 24 and 32, respectively), which are included within the medial PFC (mPFC). The mPFC also includes BA 6, 8, 9 and 10; dorsolateral PFC (DLPFC) includes BA 6, 8, 9 and 46; and the inferior frontal gyrus (IFG) and ventrolateral PFC (vlPFC) encompass inferior portions of BA 8, 44 and 45. These various processes and regions participate to a different degree in craving, intoxication, bingeing and withdrawal. lOFC, lateral OFC; mOFC, medial OFC; PFC, prefrontal cortex. Reprinted with permission from (Goldstein and Volkow 2012).

treatment of this disease. But at the same time, the fact that addiction is a brain disease that impacts the very neural fabric that enables self-monitoring, self-determination and complex social functioning indicates that a fundamental revisiting of society's conventional responses to the problem of substance use disorders is long overdue.

References

Chang L, Chronicle EP (2007): Functional imaging studies in cannabis users. *Neuroscientist* 13:422-432.

Di Chiara G (1999): Drug addiction as dopamine-dependent associative learning disorder. *Eur J Pharmacol* 375:13-30.

Ernst M, Paulus MP (2005): Neurobiology of decision making: a selective review from a neurocognitive and clinical perspective. *Biol Psychiatry* 58:597-604.

Forbes CE, Grafman J (2010): The role of the human prefrontal cortex in social cognition and moral judgment. *Annu Rev Neurosci* 33:299-324.

Goldstein R, Woicik PA, Maloney T, Tomasi D, Alia-Klein N, Shan J, et al (2010): Oral methylphenidate normalizes cingulate activity in cocaine addiction during a salient cognitive task. *Proc Natl Acad Sci U S A* 107:16667-16672.

Goldstein RZ, Volkow ND (2002): Drug addiction and its underlying neurobiological basis: neuroimaging evidence for the involvement of the frontal cortex. *Am J Psychiatry* 159:1642-1652.

Goldstein RZ, Volkow ND (2012): Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. *Nat Rev Neurosci* 12:652-669.

Kalivas PW, Volkow ND (2005): The neural basis of addiction: a pathology of motivation and choice. *Am J Psychiatry* 162:1403-1413.

Karelina K, DeVries AC (2011): Modeling social influences on human health. *Psychosom Med* 73:67-74.

Koenigs M, Young L, Adolphs R, Tranel D, Cushman F, Hauser M, et al (2007): Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446:908-911.

Kolb B, Mychasiuk R, Muhammad A, Li Y, Frost DO, Gibb R (2012): Experience and the developing prefrontal cortex. *Proc Natl Acad Sci U S A*.

McLellan AT, Skipper GS, Campbell M, DuPont RL (2008): Five year outcomes in a cohort study of physicians treated for substance use disorders in the United States. *BMJ* 337:a2038.

Seo DC, Huang Y (2012): Systematic review of social network analysis in adolescent cigarette smoking behavior. *J Sch Health* 82:21-27.

Simoni-Wastila L, Yang HK (2006): Psychoactive drug abuse in older adults. *Am J Geriatr Pharmacother* 4:380-394.

St Onge JR, Abhari H, Floresco SB (2011): Dissociable contributions by prefrontal D1 and D2 receptors to risk-based decision making. *J Neurosci* 31:8625-8633.

Volkow N, Fowler J (2000): Addiction, a disease of compulsion and drive: involvement of the orbitofrontal cortex. *Cereb Cortex* 10:318-325.

Volkow N, Fowler J, Wang GJ (2003): The addicted human brain: insights from imaging studies. *J Clin Invest* 111:1444-1451.

Volkow ND, Chang L, Wang GJ, Fowler JS, Ding YS, Sedler M, et al (2001): Low level of brain dopamine D2 receptors in methamphetamine abusers: association

with metabolism in the orbitofrontal cortex. *Am J Psychiatry* 158:2015-2021.

Volkow ND, Fowler JS, Wang GJ, Hitzemann R, Logan J, Schlyer DJ, et al (1993): Decreased dopamine D2 receptor availability is associated with reduced frontal metabolism in cocaine abusers. *Synapse* 14:169-177.

Volkow ND, Wang GJ, Begleiter H, Porjesz B, Fowler JS, Telang F, et al (2006): High levels of dopamine D2 receptors in unaffected members of alcoholic families: possible protective factors. *Arch Gen Psychiatry* 63:999-1008.

Volkow ND, Wang GJ, Fowler JS, Logan J, Gatley SJ, Hitzemann R, et al (1997): Decreased striatal dopaminergic responsiveness in detoxified cocaine-dependent subjects. *Nature* 386:830-833.

Volkow ND, Wang GJ, Telang F, Fowler JS, Logan J, Jayne M, et al (2007): Profound decreases in dopamine release in striatum in detoxified alcoholics: possible orbitofrontal involvement. *The Journal of neuroscience: the official journal of the Society for Neuroscience* 27:12700-12706.

Volkow ND, Wang GJ, Tomasi D, Baler RD (2012): Obesity and addiction: neurobiological overlaps. *Obes Rev*.

Weiss F, Koob GF (2001): Drug addiction: functional neurotoxicity of the brain reward systems. *Neurotox Res* 3:145-156.

Yuan K, Qin W, Wang G, Zeng F, Zhao L, Yang X, et al (2012): Microstructure abnormalities in adolescents with internet addiction disorder. *PLoS One* 6:e20708.

Understanding Layers: From Neuroscience to Human Responsibility

Michael S. Gazzaniga[1]

Introduction

It was 48 years ago this month when my mentor Roger Sperry spoke here on the occasion of the Pontifical Academy of Sciences symposium, *Brain and Conscious Experience*. I remember the event well as Dr. Sperry was speaking about our research on the original "split-brain" patients – studies of patients who had undergone epilepsy surgery separating the two halves of the brain. In rereading that paper it is interesting to note that the participants commented only on those studies, not on his rather extensive arguments dealing with the problem of mind and free will. This is a shame because his thoughts on free will were quite clear and indeed set the stage for many discussions since that time.

Sperry segued from discussing the patients to the issue of free will by suggesting that the research had indicated that, with the slice of a surgeon's knife, one brain might become two, each with its own set of controls. This suggestion was immediately challenged by two fellow neuroscientists, Sir John Eccles and Donald MacKay. At that conference and in the years that followed, Eccles argued that the right hemisphere had a limited kind of self-consciousness, but not enough to bestow personhood, which resided in the left hemisphere. Donald Mackay was not satisfied with the idea either and commented in his Gifford lecture some ten years later, "But I would say that the idea that you can create two individuals merely by splitting the organizing system at the level of the corpus callosum which links the cerebral hemispheres is unwarranted by any of the evidence so far ... It is also in a very important sense implausible".

[1] SAGE Center for the Study of Mind, University of California, Santa Barbara. Supported by the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The John Templeton Foundation and SAGE Center for the Study of Mind, UCSB. I would like to thank several colleagues for their insights and critiques. Steven Hillyard, Michael Posner, Danniel Bassett, Walter Sinnott-Armstrong, Jane Nevins, Rebecca Gazzaniga, Marin Gazzaniga, and Charlotte Smylie.

These concerns were all about the meaning of split-brain research, not about the issue of determinism and free will. As I have reviewed elsewhere (Gazzaniga, 2011), many early interpretations of the meaning of split-brain work have been modified, leaving this aspect of the debate moot. On the larger question of determinism and free will, Sperry's original thoughts remain clear. It is worthwhile to remind ourselves of what he said.

> Unlike *mind, consciousness,* and *instinct, free will* has not made any notable comeback in behavioral science in recent years. Most behavioral scientists will refuse to recognize the presence of free will in brain function. Every advance in the science of behavior, whether it has come from the psychiatrist's couch, from microelectrode recording, from brain splitting, from the use of psychomimetic drugs, or from the running of cannibalistic flatworms, seems only to reinforce that old suspicion that free will is just an illusion like the rise and setting of the sun. The more we study and learn about the brain and behavior, the more deterministic, lawful and causal it appears.
>
> In other words, behavioral science tells us that there is no reason to think that any of us here today had any real choice to be anywhere else, nor even to believe in principle that our presence here was not already in the cards, so to speak, five, ten or fifteen years ago. I do not like or feel comfortable about this kind of thinking any more than you do, but so far I have not found any satisfactory way around it. Alternatives to the rule of causal determinism in behavior that I have seen proposed so far, as for example, the inferred unlawfulness in the dance of subatomic particles, seem decidedly more to be deplored as a solution than desired.
>
> This is not to say that in the practice of behavioral science we have to regard the brain as just a pawn of the physical and chemical forces that play in and around it. Far from it. Recall that a molecule in many respects is the master of its inner atoms and electrons. The latter are hauled and forced about in chemical interactions by the overall configurational properties of the whole molecule. At the same time, if our given molecule is itself part of a single-celled organism like paramecium, it in turn is obliged, with all its parts and its partners, to follow along a trail of events in time and space determined largely by the extrinsic overall dynamics of *Paramecium caudatum*. And similarly, when it comes to brains, remember always that the simpler electric, atomic, molecular, and cellular forces and laws, though still present and operating, have all been superseded in brain dynamics by the configurational forces of higher level mechanisms. At the top, in the

> human brain, these include the powers of perception, cognition, memory, reason, judgment, and the like, the operational, causal effects of forces of which are equally or more potent in brain dynamics than are the outclassed inner chemical forces.
>
> You sense the underlying rationalization we are leading to here: "If you can't lick'em, join'em". If we cannot avoid determinism, accept and work with it. There may be worse "fates" than causal determinism. Maybe after all it is better to be properly imbedded in the causal flow of cosmic forces, as an integral part thereof, than to be on the loose and out of contact, free-floating, as it were, with behavioral possibilities that have no antecedent cause and hence no reason or any reliability for future plans or predictions.

Sperry captures many ideas in this his own summary of his views. Overall he articulates well that elements of all kinds become something else when configurational issues are accounted for. Some kind of other complexity arises out of interacting parts and that new layer can constrain the very elements that produced it.

Still, this formulation was not widely accepted at the time and other participants at the conference in their own contributions contested Sperry's view and for markedly different reasons. Again, Eccles and Mackay challenged his ideas. Eccles was a dualist and believed the mental inserted itself into the brain in the left supplementary motor area. Mackay, on the other hand, agreed that the brain "was as mechanical as clockwork". However, he believed there was what he called a "logical indeterminacy" that kept free will alive. This was the concept that in order for something to be true it had to be true for everybody at all times. Thus if a super brain scientist made a prediction about my future actions, all I would have to do to negate the prediction is not carry out the act at a prescribed time. If the super brain scientist wrote down the prediction and sealed it in an envelope and sure enough I did what he predicted, that prediction still wouldn't count since for something to be true and valid, it has to be known to all at all times. Once it is known to all, the person can choose not to carry out the prediction and so on.

While this debate raged on for years, it was somewhat local to neuroscience. The philosophers, by and large, were and still are coming at the problem from different angles, far too many to review here. What is relevant to the current effort is the strong belief among philosophers that it is difficult to separate the issue of free will from the issue of responsibility. This traditional position, which is well represented by Daniel Dennett, finds people viewing the determined brain as in fact an exemplar of a "free" system.

As Dennett has recently written:

> When, on the other hand, we have our wits about us, and are not massively misinformed or otherwise manipulated, then *there is no important sense* (emphasis in the original) in which the outcome of all the interactions in the many levels or layers of "machinery" is not a free choice. That's what a free choice *is!* It's the undistorted, unhindered outcome of a cognitive/conative/emotive process of exquisite subtlety, capable of canvassing the options with good judgment and then acting with fairly full knowledge of what is at stake and what is likely to transpire. (Dennett, 2013)

With this kind of definition of what it means to be "free" the yoking of free will and responsibility remains intact. While other philosophers don't see this need and indeed claim the two concepts are dissociable (Fischer and Ravissa, 1999), it is a key issue. In short, does a deterministic view of brain function make nonsense out of the idea of responsibility?

I join with those who believe one can hold a deterministic view and still maintain humans are personally responsible for their actions. In what follows I will make this argument by suggesting a layered view of human decision making that incorporates the social network within which we live, thereby making the idea not only plausible but inevitable and necessary. It is this perspective that advances the ideas of Sperry's contribution almost 50 years ago. While incorporating the mental realm in the causal flow is important as Sperry noted, it does not liberate one from the reality of determinism. After all, the mental layer, with its own abstract vocabulary and mechanisms, works in a deterministic way as well.

By recognizing the existence of yet another layer, the social layer, another level of abstraction and explanation is introduced which does impose a constraint on the individual. Being accountable and therefore responsible is the *sine qua non* of existing in a social world. In short, responsibility is established by participating in the social network; it is not found in the brain per se.

Towards Layered and Dynamical Views of Brain/Mind Function

Much of what follows, I have presented elsewhere (Gazzaniga, 2011, 2013; Bassett and Gazzaniga, 2011). In those efforts, I reviewed neuroscientific data which supports the modular view of brain organization, now widely established, along with a possible understanding of why our subjective life seems largely unified.

From today's vantage point: It's all about the brain – what it does and does not do. First, how is that thing built and connected, and how does it work? Is it a bowl of mush shaped by its environment, like a wheelbarrow

full of wet concrete being poured into a form? Or does the brain arrive on the scene pre-formed, to some extent, and then await experience to place the final touches on its mature shape? More importantly, does it matter how it is built, for the purposes of this discussion?

It does. We are born with an intricate brain slowly developing under genetic control, with refinements being made under the influence of epigenetic factors and activity-dependent learning. It displays structured, not random, complexity, with automatic processing, with particular skill sets, with constraints, and with a capacity to generalize. All these evolved through natural selection and provide the foundation for a myriad of cognitive abilities that are separated and represented in different parts of the brain. These parts feature distinct but interrelated neural networks and systems. In short, the brain has distributed systems running simultaneously and in parallel. It has multiple control systems, not just one. Our personal narrative comes from this brain, and how it interprets the outside world within which it lives.

This overall neural architecture has been unearthed at many levels of examination. While developmental neurobiologists have revealed how the brain gets built, cognitive neuroscientists have studied the brain in healthy maturity and often when it is damaged. My colleagues and I used those insights to confirm that there are modularized, and frequently localized, processes in the functioning, fully developed brain. Classic studies on neurologic patients by Broca and others supported the idea that brain injury can lead to the loss of specific cognitive abilities. This notion has been the backbone of behavioral neurology. Split-brain research complemented this work. It showed what happened when one processing system was disconnected from others, even though it was still present and functioning. And what did happen? It just went on functioning, outside the realm of awareness of the other systems. The right brain was able to go about its business normally while the left brain didn't have the slightest idea what the right brain was doing – and vice-versa.

Still, this emerging knowledge of how our brain is organized was hard to square with ordinary experience. People – even split-brain patients – feel integrated, whole and purposeful, not modularized and multiple. How can our sense of being singular and responsible come from a neural architecture like ours?

The Interpreter of Experience

Years ago we unearthed a special capacity, a module in the left hemisphere that we called the "interpreter". Studies of split-brain patients demonstrated that each side of the brain could respond separately to queries about what it perceives by having the hand it controls point to answers in

a multiple-choice task. So flash a picture of a chicken claw to the left brain, and the right hand could choose a picture of a chicken out of a group of pictures (each side of the brain controls the opposite side of the body). If the right brain was at the same time shown a picture of a snow scene, it could guide the left hand to select a picture of a snow shovel from a different set of pictures. It took us years to figure out the key question to ask after a split brain patient performed this task: "Why did you do that?"

We arranged for one patient's left hemisphere (which controls speech) to watch the left and right hand pointing to two different pictures while not allowing the left brain to see the snow scene. Of course, the left hemisphere knew why the hand it controlled had pointed to the chicken, but it had no access to information about why the patient's left hand, controlled by the right hemisphere, had pointed to the shovel. Nonetheless, immediately upon being asked our key question, "why did you do that?", the left hemisphere made up a story, an interpretation, of why the left hand, controlled by a separated brain module, did what it did. The patient answered, "Oh, the chicken claw goes with the chicken and you need a shovel to clean out the chicken shed".

Years of research have confirmed that there is a system that builds a narrative in each of us about why we do things we do, even though our behaviors are the product of a highly modularized and automatic brain working at several different layers of function (Gazzaniga, 2000). Our dispositions, quick emotional reactions, and past learned behavior are all fodder for the interpreter to observe. The interpreter finds causes and builds our story, our sense of self. It asks, for example, "Who is in charge?" and in the end concludes, "Well, it looks like I am".

Additionally, neuroscientists have continued to examine when the brain carries out its work that is associated with behavior or even conscious activity itself. Ever since the classic work of Benjamin Libet, it has been believed that the neural events associated with an action occur long before one is consciously aware of even wanting to will an act. Libet stimulated the brain of an awake patient during the course of a neurosurgical procedure and found that there was a time lapse between the stimulation of the cortical surface that represents the hand and when the patient was conscious of the sensation in the hand (Libet *et al.*, 1979). In later experiments, brain activity involved in the initiation of an action (pushing a button), occurred about five hundred milliseconds *before* the action. What was surprising was that there was increasing brain activity related to the action as many as three hundred milliseconds *before* the *conscious intention* to act according to subject reports. The buildup of electrical charge within the brain that preceded

what were considered conscious decisions was called Bereitschafts potential or more simply, the readiness potential (Libet *et al.*, 1983). Using more sophisticated fMRI techniques, John-Dylan Haynes (Soon *et al.*, 2008) recently showed that the outcomes of an inclination can be encoded in brain activity up to ten seconds before it enters awareness! Furthermore, the brain scan can be used to make a prediction about what the person is going to do. The implications of this result appear definitive. They suggest completes its work independent of conscious input.

These sorts of findings, however, can be interpreted differently when the brain is viewed as a multi-layered system as is commonly seen in information systems (see Hillis, 1998; Bachman *et al.*, 2000; also see Doyle and Csete, 2011). Simply put, layered systems use layers to separate different units of functionality. Each layer preferentially communicates with the layer above and the layer below. Each layer *uses* the layer below to perform its function… "A **Layer** is a design construct. It is implemented by any number of classes or modules that behave like they are all in the same layer. That means that they only communicate with classes in layers immediately above or below their layer and with themselves" (Van Bergen, P.).

The framing of how the brain manages its tasks will undoubtedly be modified and extended in the years to come. Still, this suggested informational/functional assessment of how the brain does its work is liberating as it frees us from the linear assumptions of bottom-up causality. The traditional reductionist/constructionist approach with its claim of linearity on how the brain produces mental states leaves little apparent room for the role of mental life in human destiny. On the surface that seems absurd. This was Sperry's point 50 years ago and it is as valid today as it was then.

Clearly, we humans enjoy mental states that arise from our underlying neuronal, cell-to-cell interactions. Mental states do not exist without those interactions. However, as argued in the foregoing, mental states cannot be defined or understood by knowing only the cellular interactions. Mental states that emerge from our neural actions, do constrain the very brain activity that gave rise to them, just as Sperry noted that "a molecule in many respects is the master of its inner atoms and electrons". Mental states, such as beliefs, thoughts, and desires, represent a layer, and that layer arises from brain activity and in turn can and does influence our decisions to act one way or another. Ultimately, these interactions will be understood only with a new vocabulary that captures the fact that two different layers of stuff are interacting in such a way that existing alone animates neither.

Yet, this interpretation of the problem, where both upward and downward causation are discussed, comes with warning signs. As John Doyle puts

the issue (See Gazzaniga, 2011):

> ... the standard problem is illustrated with hardware and software; software depends on hardware to work, but is also in some sense more 'fundamental' in that it is what delivers function. So what causes what? Nothing is mysterious here, but using the language of 'cause' seems to muddle it. We should probably come up with new and appropriate language rather than try to get into some Aristotelian categories.

Understanding this nexus and finding the right language to describe it represents, as Doyle says, "the hardest and most unique problem in science" (Personal Communication). The freedom represented in a choice not to eat the jelly donut comes from a mental-layer belief about health and weight, and it can trump the pull to eat the donut because a certain brain module likes its taste. The bottom-up pull sometimes loses out to a top-down belief in the battle to initiate an action. And yet the top layer does not function alone or without the participation of the bottom layer.

A unique vocabulary which has yet to be developed, is needed to capture the thing that happens when mental processes constrain the brain and vice versa. The action is at the interface between those layers. In one kind of vocabulary, it is where downward causation meets upward causation. In still another perspective, it is not only there but also in the space between brains that are interacting with each other. Overall, what happens at the interface of our layered hierarchical existence holds the answer to our quest for understanding mind/brain relationships. How are we to describe that? Recalling Libet and Haynes, we have to account for the role of time. I think we should say that mind/brain layers interacting has its own time course and that time course is current with the actions taking place. In short, it is the abstract interactions between the mind/brain layers that make us current in time, real and accountable to our past mental experiences. The whole business about the brain doing it before we are conscious of it becomes moot and inconsequential from the vantage point of a layered interacting system. Again and as I have discussed elsewhere:

> Once a mental state exists, is there downward causation? Can a thought constrain the very brain that produced it? Does the whole constrain its parts? This is the 64 thousand dollar question in this business. The classic puzzle is usually put this way: There is a physical state, P1, at time 1, which produces a mental state, M1. Then after a bit of time, now time 2, there is another physical state, P2, which produces another mental state, M2. How do we get from M1 to M2? This is the conundrum. We know that mental states are produced from processes in the brain so that M1 does not directly generate M2 with-

out involving the brain. If we just go from P1 to P2 then to M2, then our mental life is doing no work and we are truly just along for the ride. No one really likes that notion. The tough question is, does M1, in some downward constraining process guide P2, thus affecting M2? We may get a little help with this question from the geneticists. They used to think gene replication was a simple upwardly causal system: Genes were like beads on a string that make up a chromosome that replicates and produces identical copies of itself. Now, they know that genes are not that simple, there is a multiplicity of events going on. Our systems-control guy, Howard Pattee, finds that a good example of upward and downward causation is the genotype-phenotype mapping of description to construction. It "requires the gene to describe the sequence of parts forming enzymes, and that description, in turn, requires the enzymes to read the description… In its simplest logical form, the parts represented by symbols (codons) are, in part, controlling the construction of the whole (enzymes), but the whole is, in part, controlling the identification of the parts (translation) and the construction itself (protein synthesis)". And once again Pattee wags his finger at extreme positions that champion which is more important, upward or downward. As a teenager would sum it up, "Duh, they are like, complementary".

It is this sort of analysis that finds me realizing the reasoning trap we can all too easily fall into when we look to the Libet kind of fact, that the brain does something before we are consciously aware of it. With the arrow of time all moving in one direction, with the notion that everything is caused by something before it, we lose a grip on the concept of complementarities. What difference does it make if brain activity goes on before we are consciously aware of something? Consciousness is its own abstraction on its own time scale and that time scale is current with respect to it. Thus, the Libet thinking is wrong headed. That is not where the action is anymore than a transistor is where the software action is.

Setting a course of action is automatic, deterministic, modularized and driven not by one physical system at any one time but by hundreds, thousand and perhaps millions. The course of action taken appears to us as a matter of "choice" but the fact is, it is the result of a particular emergent mental state being selected by the complex interacting surrounding milieu. Action is made up of complementary components arising from within and without. That is how the machine (brain) works. Thus, the idea of downward causation might confuse our un-

> derstanding. As John Doyle says, "where is the cause?" What is going on is the match between ever present multiple mental states and the impinging contextual forces within which it functions. Our interpreter then claims we freely made a choice. (Gazzaniga, 2011)

It is also true that viewing the brain/mind interface from this perspective reveals a certain truth: the brain is a dynamical system. Instead of working in a simple linear way where one thing produces another, it works in a dynamic way where two layers interact to produce a function. Hardware and software interact to produce the PowerPoint image. Mental states interact with neuronal states to produce conscious states. Starting the clock on what happens when, when trying to analyze the flow of events during conscious activity, doesn't start with neurons firing off, as those events might reflect little more than the brain warming up for its participation in the dynamic events. The time line starts at the moment of the interaction between layers. At the level of human experience, that would mean we are all online when we are thinking about whatever we are thinking about. Thought is not on a delay after action. It also leads to the question of whether or not mental beliefs can be in the flow of events determining ultimate action (Posner and Rothbart, 2012). I think so.

Moving Forward: Emergence, Human Responsibility and Freedom

In one sense, the concept of multiple levels has a long-standing history in the study of the brain and mind. For literally thousands of years, philosophers have argued about whether the mind and body are one entity or two. The compelling idea that people are more than just a body, that there is an essence, a spirit or mind, has been around a long time. What has not been fully appreciated, however, is that viewing the mind/brain system as a layered system sets the stage for understanding how the system actually works. As reviewed in the foregoing pages, it also allows for understanding the role of how beliefs and mental states stay part of our determined system. With that understanding comes the insight that layers exist both below the mind/brain layers and above them as well. Indeed, there is a social layer and it is in the context of interactions with that layer that we can begin to understand concepts such as personal responsibility.

I believe that we neuroscientists are looking at the concept of responsibility at the wrong organizational level. Put simply, we are examining it at the level of the individual brain when perhaps responsibility is a property of social groups of many brains interacting. Mario Bunge makes the point that we neuroscientists should heed: "… we must place the thing of interest in its context instead of treating it as a solitary individual". By placing such

concepts as personal responsibility in the social layer, it removes us from the quagmire of understanding how determined brain states negatively influence responsibility for our actions. Being personally responsible is a social rule of a group, not a mechanism of a single brain.

Sperry did not introduce the idea of the social layer. He fully accepted and indeed implored us all to "join'em if you can't lick'em". He did feel better about determinism by conceptualizing it was the mind layer that intervenes and becomes a part of the causal chain that constrains the neural elements that built the mind. With the present view, adding the social layer to the human condition completely restores the idea of personal and therefore moral responsibility no matter how stringent a deterministic stance one adopts. Responsibility comes out of the agreement humans have with each other to live in the social world. The human social network is like any other kind of network. The participants have to be held accountable for their actions—their participation. Without that rule, nothing works.

Brains are automatic machines following hierarchical decision pathways and analyzing single brains in isolation cannot illuminate the capacity to be responsible. Again, responsibility is a dimension of life that comes from social exchange, and social exchange requires more than one brain. When more than one brain interacts, a new set of rules comes into play and new properties – such as personal responsibility – begin to emerge. The properties of responsibility are found in the space between brains, in the interactions between people.

Finally, neuroscience is happy to accept that human behavior is the product of a determined system, which is guided by experience. But does it matter how that experience is doing the guiding? If the brain is a decision-making device and gathers information to inform those decisions, then can a mental state that is the result of some experience or the result of some social interaction affect or constrain the brain and with it future mental states?

We humans are about becoming less dumb, about making better decisions to cope and adapt to the world we live in. That is what our brain is for and what it does. It makes decisions based on experience, innate biases, and much more. Our "freedom" is to be found in developing more options for our computing brains to choose between. As we move though time and space we are constantly generating new thoughts, ideas, and beliefs. All of these mental states provide a rich array of possible actions for us. The couch potato simply does not have the same array as the explorer. Just as Daniel Dennett suggests, even though we live in a determined world, new experience provides the window into more choices and that is what freedom truly means.

Personal responsibility is another matter. My argument is that it is real, the consequence of social strategies that people adopt when living together and that are the fabric of social life. Personal responsibility is not to be found in the brain, any more than traffic can be understood by knowing about everything inside a car. I am inclined to think there is something like a universal architectural principle common to all information processing systems. All networks, whether they are neural or artefactual like the Internet, can operate only if accountability – cause and effect, action and consequence – is built into their functioning. Human society is the same.

It could be argued that the addition of the social layer actually reduces rather than enhances personal responsibility – responsibility in the sense of an individual having a choice about their actions. If the system is deterministic and the social layer impinges from above along with the mental layer while the brain impinges from below it would seem that the individual (wherever he or she is located amongst the layers) doesn't have much choice about what to do. His or her actions are determined, but not by him or her alone, but by the interactions of these various networks. "The network made me do it!"

This line of reasoning slips into place the assumption there is a control center, indeed a homunculus, a thing that is calling the shots. What modern neuroscience perspectives argue for, however, is that the brain is an information processing system whose function it is make decisions for actions. In carrying out that function it gathers information from multiple layers with the impact of each layer being evaluated by the neural algorithms that manage choice for action. Framing the problem in this way removes the endlessly circular problem of defining the self with its seemingly appealing solution to providing an explanation for the mind's need to have an essential mechanism in charge. As I have summarized elsewhere:

> Understanding that the brain works automatically and follows the laws of the natural world is both heartening and revealing. Heartening because we can be confident the decision-making device, the brain, has a reliable structure in place to execute decisions for actions. It is also revealing, because it makes clear that the whole arcane issue about free will is a miscast concept, based on social and psychological beliefs held at particular times in human history that have not been borne out and/or are at odds with modern scientific knowledge about the nature of our universe. As (Caltech's) John Doyle has put it to me, "Somehow we got used to the idea that when a system appears to exhibit coherent, integrated function and behavior, there must be some 'essential' and, importantly, central or centralized controlling element that is responsible. We are deeply essentialist, and our left brain

> will find it. And as you point out, we'll make up something if we can't find it. We call it a homunculus, mind, soul, gene, etc. … But it is rarely there in the usual reductionist sense … that doesn't mean there isn't in fact some 'essence' that is responsible, it's just distributed. It's in the protocols, the rules, the algorithms, the software. It's how cells, ant hills, Internets, armies, brains, really work. It's difficult for us because it doesn't reside in some box somewhere, indeed it would be a design flaw if it did because that box would be a single point of failure. It's, in fact, important that it not be in the modules but in the rules that they must obey".

Again, the idea here is that the whole concept of personal responsibility is a social concept, automatically assigned and pinned on the individual by the group de facto. It is a fact of the social world, the price of entry and participation. It is part of the architecture of human existence.

Responsibility is a needed consequence of more than one individual interacting with another. It is established by people. Researchers might study the mechanistic ways of the brain-mind interface forever, with each year yielding more insights. Yet *all* of their research can only add to the scientific case for the central value of human life. It is because we have a contract with our social milieu, we are responsible for our actions.

References

Anderson, P.A. (1972). More is different. *Science*, *177*(4047), 393-396.

Bachman, F., Bass, L., Clements, P., Garlan, D., Ivers, J., Little, R., Nord, R and Stafford, J. (2000) *Documenting Software Architecture: Documenting Interfaces. Technical note*. Carnegie Mellon, Software Engineering Institute.

Bassett DS, & Gazzaniga MS. Understanding complexity in the human brain, *Trends Cogn Sci*. 2011 May; 15(5):200-9. Epub 2011 Apr 14.

Bohr, M. (1937) Causality and Complementarity. *Philosophy of Science*, *4*(3), 289-298.

Dennett, D. (2013) Seduced by Tradition: Comment on Gazzaniga's paper. In: *Moral Psychology*, Volume 4: Free Will and Moral Responsibility, Ed. Walter Sinnott-Armstrong Cambridge, Mass.; MIT Press, 2013.

Doyle, John (2011) personal communication.

Doyle JC & Csete ME. Architecture, Constraints, and Behavior, *P Natl Acad Sci* USA, (2011) vol. 108, Sup 15624-15630.

Eccles, John C. (1980) *The Human Psyche* (Gifford Lectures) Springer-Verlag.

Fischer, John Martin, & Revissa, Mark (1999) *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Gazzaniga, M.S. (2005) *The Ethical Brain*. New York: Harper Perennial.

Gazzaniga, M.S. *Human – The Science Behind What Makes Us Unique*. Ecco Books, Harper, New York 2008.

Gazzaniga, Michael S., Neuroscience and the Correct Level of Explanation For Understanding Mind 14, *Trends in Cognitive Sciences* 291 (2010)

Gazzaniga, Michael S. (2011) Who's in

Charge, Free Will and the Science of the Brain, Harper (Ecco) New York.

Gazzaniga, M.S. (2013) Mental Life And Responsibility In Real Time With A Determined Brain. In: *Moral Psychology*, Volume 4: Free Will and Moral Responsibility, Ed. Walter Sinnott-Armstrong Cambridge, Mass.; MIT Press, 2013.

Hillis, W. Daniel (1999) *The Pattern On The Stone: The Simple Ideas That Make Computers Work* (Science Masters) Basic Books, New York.

Libet, B., Gleason, C.A., Wright, E.W., & Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious initiation of a freely voluntary act. *Brain*, 106 (3), 623-642.

Libet, B., Wright, E.W., Feinstein, B. & Pearl, D.K. (1979). Subjective referral of the timing for a conscious sensory experience: A functional role for the somatosensory specific projection system in man. *Brain*, 102(1), 193-224.

MacKay, D.M. (1991) *Behind the eye*. Oxford: Basil Blackwell.

Posner, Michael & Rothbart, M. (2012) (in Press) Willpower and Brain Networks, *Bulletin of the International Society for the Study of Behavioural Development* (ISSBD) Special issue on Neuroscience and Development.

Soon, C.S., Brass, M., Heinze, H.-J., Haynes, J.-D. (2008) Unconscious determinants of free decision in the human brain. *Nature Neuroscience*, 11(5), 543-545.

Sperry, R.W. (1966) Brain bisection and mechanisms of consciousness. In: J.C. Eccles (Ed.), Brain and Conscious Experience, pp. 298-313. Heidelberg: Springer-Verlag.

Van Bergen, P. online at www.dossier-andreas.net/software_architecture/index.html

Self-Knowledge and the Adaptive Unconscious

Timothy D. Wilson[1]

At the dawn of human consciousness, when people first gained the ability to reflect upon the world around them, it seems likely that they turned the spotlight of consciousness inward to try to understand themselves. After all, throughout recorded human history, self-knowledge has been a highly valued trait. For thousands of years, Buddhists have sought greater self-awareness through the practice of meditation. The Greeks inscribed "Know Thyself" on the wall of the temple at Delphi. Some of the most respected figures in the Catholic Church extolled the virtue of self-knowledge, such as St Augustine of Hippo, who in his prayer for self-knowledge wrote, "Lord Jesus, let me know myself and know You" (Augustinian Spirituality, n.d.). Similarly, St Teresa of Avila suggested that, "Self-knowledge is of such consequence that I would not have you careless of it" (St Teresa of Avila, 1577).

To be sure, different religions and philosophical approaches emphasize different aspects of self-knowledge. In Buddhism it is a realization of the transitory nature and unimportance of the self, and a route by which people can gain awareness of suffering and compassion toward others, as well as a greater awareness of one's feelings and motives (Flanagan, 2011). In Catholicism, it is gaining a sense of humility and an appreciation of God's power. For example, St Augustine's prayer goes on to say, "Let me banish self and follow You, and ever desire to follow You" (Augustinian Spirituality, n.d.), and St Teresa of Avila adds, "I believe we shall never learn to know ourselves except by endeavouring to know God, for, beholding His greatness we are struck by our own baseness" (St Teresa of Avila, 1577).

Psychological science shares the idea that self-knowledge is of paramount importance, but with a somewhat different focus than religious teachings. Rather than spiritual growth or a sense of humility, psychology has focused on the value of self-knowledge (e.g., to mental health) and how people attain it. Here I will review research on each of these topics.

[1] The preparation of this paper was aided by a grant from the National Science Foundation Grant SES-0951779. Correspondence concerning this article should be sent to Timothy D. Wilson, Department of Psychology, University of Virginia, P.O. Box 400400, Charlottesville, VA 22904-4400. E-mail: tdw@virginia.edu.

Is There Value to Self-Knowledge?

The simple answer to this question is, "Of course". People who are completely out of touch with their abilities, traits, and feelings are likely to be unhappy with themselves and insufferable around others. Indeed, a common definition of mental illness is a loss of touch with reality, including one's own traits and capabilities. A person who believes he can fly after jumping off of tall buildings is unlikely to live for very long, and a teenager who is convinced that she will be the next star on the *American Idol* television show, despite an inability to carry a tune, is destined for failure and heartache.

But the question of how valuable self-knowledge is turns out to be more complicated. Taylor and Brown (1988), in a seminal article, suggested that "positive illusions" about oneself can be beneficial. Consider two people who are both talented singers. The one who has greater faith in her ability, and perhaps even exaggerates it a bit, is likely to work harder and persevere more in the face of failure, than the one who has a more realistic view. As long as positive illusions aren't too extreme, this argument goes, they confer motivational and affective benefits.

The Taylor and Brown (1988) argument caused a good deal of controversy that continues to this day (see Vazire & Wilson, 2012). But a general consensus has emerged that positive illusions (sometimes called "self-enhancement") can be beneficial or costly, depending on a number of circumstances. One moderator, as mentioned, is how extreme the illusions are. We have a name for people who exaggerate their own talents and achievements to extreme degrees – narcissists – and research shows that such individuals pay a price for their self-enhancement: others view them as pompous and unlikeable (e.g., Colvin, Block, & Funder, 1995; Schriber & Robins, 2012).

There is also evidence that it is best is to keep our self-enhancing tendency to ourselves, because wearing them on our sleeves annoys others. Support for this hypothesis comes from a study by Dufner (2012), who found that there was a social benefit to *actual* self-enhancement but a social cost to *perceived* self-enhancement. The more people actually self-enhanced – that is, the extent to which they thought that they were more intelligent than they actually were – the more they were liked by their friends, perhaps because they were confident and happy people. But the more people were perceived to be self-enhancers by their friends, the more unpopular they were. The moral seems to be that it is good to privately believe that we are better than we are, but not to convey that view to others.

It is also important to consider the value of self-knowledge from a broader perspective. In addition to overestimating specific abilities, such as how well we can sing or how intelligent we are, are there key myths about

ourselves and the social world that increase our well-being? I believe that there are a small number of essential myths that human beings share. Whether they are adaptive is open to debate, though I suspect they are, at least in moderation. I offer four candidates of such myths here. Warning: Discussing these myths is tantamount to dispelling them, at least temporarily, so read on at your own risk!

Essential Myth 1: We Are Immortal

Everyone knows that they are mortal beings who have an expiration date. Some people believe in the immortality of the soul or in reincarnation, of course, but no one can deny that our existence in our current bodies will end sooner or later. Did you experience a ping of anxiety when you read that sentence? I confess that I experienced such a ping when I wrote it. Knowledge of our own mortality is the ultimate existential threat and people have developed all sorts of strategies to avoid thinking about the inevitability of their deaths (Solomon, Greenberg, & Pyszczynski, 2004). As a result, we live much of our lives avoiding the knowledge of our ultimate demise, or at least avoiding to think about it. Is such denial adaptive? The answer is no if it leads to behaviors that will hasten our demise, such as smoking cigarettes and overeating. But constant reminders of our ultimate end are paralyzing and I suspect that there is a happy medium, where it is best to live each hour without dwelling too much on the fact that it might be our last one, while at the same time trying to maximize our number of hours by adopting a healthy lifestyle.

Essential Myth 2: We Are Important

There is no one that we spend more time with than ourselves. As a result, it is hard to avoid the impression that we are an important force in the world, individuals who make a difference and are of great consequence to many other people. For most of us, this impression is probably not as true as we think it is. Suppose, for example, that I gave a questionnaire to your friends and family that asked them to (a) rank how important you are to them, in comparison to their other loved ones, and (b) to keep track of how often they think about you when you are not around. Next I ask you to guess what your average "importance ranking" is among your friends and family, and to guess how often you are in their thoughts. To my knowledge such a study has not been done, but if it were, I would bet that most of us would overestimate our importance and salience to others.

This myth probably helps us get out of bed in the morning, work hard at our jobs, and strike up conversations with strangers at parties. After all,

why do any of these things if we are just one insignificant speck in the universe? Of course, this myth can be taken too far, resulting in narcissism. And showing people that they are not as important as they think they are might have the benefits of reducing the number of inane posts they make on social media sites, shorten the length of their boring stories at parties, and increase the likelihood that they focus less on themselves and more on helping others. In this regard, it is interesting to note that many religions stress the insignificance of any one of us on earth, stressing humbleness over self-importance. But again, exaggerating our importance a tad probably has motivational benefits.

Essential Myth 3: The World Is as We See It

When we observe the social world and form impressions of other people, we are often surprised to learn that other people saw things differently from us. The reason we are surprised is because of a pervasive phenomenon called naïve realism, which is the assumption that we observe the world as it actually is, rather than interpreting, construing, or selecting the information that reaches our senses (Ross & Ward, 1996). Because we believe that we see things as they are, when others disagree with us, we believe that it can only be because they are wrong and we are right.

Naïve realism is not simply a motivational strategy that people adopt in order to feel good; rather, it results from the fact that people are not consciously aware of the mental processes that select and interpret information as it hits our senses. Because we cannot directly observe this process of construal, our interpretations of the world appear to us simple observations (Pronin, Gilovich, & Ross, 2004). It is also clear that naïve realism has many negative consequences. For example, it is a roadblock to resolving conflicts between adversaries; often, the two sides cannot even agree on the facts, given that each side believes that it is viewing them accurately while the other is twisting the facts to suit its own purposes. But is naïve realism also adaptive in some way? In small doses it may have some benefits. People who are never sure what the world is really like, and are constantly aware that theirs is but one of hundreds of interpretations, are likely to spend more time equivocating than acting.

Essential Myth 4: Other People are Predictable

Think of a close friend that you know really well. Now suppose that your friend found a $20 dollar bill on the floor of a bookstore. Would he or she pocket the money or try to find the person who lost it? How confident are you in your guess? Research shows that we are overconfident

when making such predictions – we are not as accurate as we think we are (e.g., Dunning, Griffin, Milojovic, & Ross, 1990). One reason for overconfidence in predicting other people's behavior is naïve realism (e.g., the assumption that our interpretation of the lost money situation is the same as how our friend would interpret it). Another is a phenomenon called the fundamental attribution error, which is the tendency to attribute other people's behavior to stable personality traits and to underestimate the role of situational factors in influencing behavior (Jones, 1979; Ross, 1977). We might be sure that our friend would return the money because we believe that he or she is an honest person. But we would likely be overestimating the extent to which honesty is the sole determinant of behavior in this situation, and underestimate the role of such situational factors as how much of a hurry our friend is in and how crowded the store is.

Are overconfident predictions adaptive in some way? As with the other myths we have discussed, it may be helpful in small doses. Obviously we don't want to be so overconfident that we are constantly surprised by what other people do. But it might be to our advantage to exaggerate our ability to predict, to some small degree. It is reassuring to believe that we can predict what our loved ones will do tomorrow, next month, and next year, rather than knowing that they might well surprise us in unsettling ways.

In sum, it is likely that the Goldilocks principle applies to each of the four myths I have just reviewed: We don't want too little or too much of them, but a level that is just right. The happiest, most effective people are probably those who don't dwell too much on their mortality, overestimate their importance to others just a little, think that they are astute observers of the world and yet can also appreciate others' points of view, and believe that they are excellent predictors of what their friends will do, while also recognizing that they are not always right.

How Do People Attain Self-Knowledge?

Although there may be benefits to some forms of positive illusions, the fact remains that it is critical to have some awareness of our own abilities, preferences, and traits. If we were clueless about what we liked, or where our talents lie, it would be difficult to accomplish much or achieve happiness. As Tennyson noted, "Self-reverence, self-knowledge, self-control – these three alone lead life to sovereign power". How, then, should we go about discovering ourselves?

Throughout most of human history, the path to self-knowledge has been thought to lie inward. Via introspection, prayer, or meditation, people focus inwardly on their own thoughts and feelings. And yet, in the past few

decades, psychological research has demonstrated the limits of introspection. As powerful as human consciousness is, it is not a powerful beam that, when turned inward, illuminates all of our motives and traits.

Sigmund Freud, of course, was one of the first to point out the limits of the conscious mind. Consciousness is the mere tip of the mental iceberg, he argued, such that most of our mental lives, including our deepest wishes, motives, and feelings, are hidden from view. And, we devote quite a bit of psychic energy to keeping them hidden, because it would be upsetting to acknowledge to ourselves how base and animal-like many of our urges are.

The modern view of the unconscious is different. While not denying the existence of the Freudian unconscious, the emphasis is more on the extent to which the everyday operations of the mind are unavailable to introspection. There is a vast *adaptive unconscious*, according to this view, that evolved over thousands of years, and includes most of the way in which the mind works (Wilson, 2002). Consciousness evolved much later, and although it is an amazing evolutionary achievement, it is not fully integrated with the rest of the mind. Put differently, human beings were probably sophisticated information processors before the dawn of consciousness, with the ability to learn, interpret incoming information, set goals, and form preferences. We have retained the ability to perform these mental feats unconsciously, without the involvement of consciousness. As a consequence, consciousness is limited not only by what we don't *want* to know, as Freud argued, but also by what we *can't* know. This is not to say that everything is unconscious. People are obviously aware of a rich set of thoughts and feelings – especially feelings, which seem to have priority in consciousness. But the point is that there is much more to the mind than what we can be aware of (Wilson, 2002).

People as Story Tellers

That doesn't mean we can't figure it out. Just as we are astute observers of other people, and make good inferences about who they are and what they are likely to do next, so can we be astute observers of ourselves. After all, we have a great deal of information to go by, more than we have about anyone else – knowledge of how we have acted in countless situations, how other people react to us, and how we feel (because, as mentioned, emotions seem especially likely to rise up into consciousness). In addition to this "raw material", we can try to see ourselves through the eyes of others, comparing our view to theirs and revising our views accordingly (Cooley, 1902; Mead, 1934; Vazire & Carlson, 2010).

In my view, self-knowledge is all about the way we integrate this raw material into stories about ourselves. Like narrative interpretations of a text,

there is not a single "true" story. Narratives are arbitrary to a degree; given the same "raw data" about others and ourselves, we can arrive at radically different interpretations. A person could view her spouse as the love of her life who should be forgiven his minor foibles and flaws, such as the fact that he once had an affair and was not the best father in the world. Or, she could weave a quite different story, ruing the day she met such a cad. Similarly, there is considerable latitude on the narratives that we construct about ourselves. Based on the same "data", a woman could view herself as a talented professional who, despite some setbacks, has become a star in her field, or as an imposter who has succeeded through blind luck and will be drummed out of her discipline as soon as people catch on.

This raises the important question of how we can judge the "goodness" of a self-story. At one extreme, postmodern psychologists have argued that no story is truer than another, and that we should avoid accuracy as a criterion (e.g., Gergen & Kaye, 1992). But surely this argument goes too far. Even postmodernists would agree, I presume, that it is not adaptive for people to believe that if they jumped off of a tall building they could fly. Indeed, research shows that the more people's conscious beliefs about their goals match their nonconscious goals, the happier they are. Consider, for example, two people who are deciding on a career. One believes that he is a "people person" and thus chooses a career that involves a lot of social interaction, such as sales. In fact, though, he has a low implicit need for affiliation, and is unlikely to be happy or successful in career that is mismatched to his implicit goals. The other person knows that she is not a "people person", and chooses a career more suited to her goals, such as becoming an accountant. Research shows that the latter person will be happier than the former (Schultheiss & Strasser, 2012).

Thus, a "good" self-story should capture who we are, at least to some extent – and by that I mean that it should correspond to our unconscious traits, needs, and goals. This doesn't mean that there is only way of telling that story, however. Indeed, late in his life, even Freud came to view the process of psychotherapy not as "uncovering truths", but rather the construction of a narrative that provides people with healthy, coherent explanations of who they are (Spence, 1982). But regardless of how the story is told, it should correspond to the person's adaptive unconscious.

Which brings us to another sign of a good self-story, namely what I have called the "peace of mind" criterion (Wilson, 2002). Our stories should provide us with a meaningful narrative that allows us to gain closure on negative episodes in our lives, instead of ruminating about them (Wilson & Gilbert, 2008). Writing exercises have been developed to help people find

meaning in traumatic events, which, in the current terminology, allows them to revise their stories in ways that make sense of their lives and allow them to move on (Pennebaker, 2004; Kross, 2009).

Finally, good self-stories should meet a "believability" criterion. Regardless of how true the story is, it should be convincing to the one who tells it. One wonders how dedicated postmodernists get through their day, if they really believe that the self-story they have adopted at noon is completely arbitrary and likely to be different from the one they adopt by dinner time. As Freud noted, in discussing the process of psychoanalysis, "an assured conviction of the truth of the construction ... achieves the same therapeutic result as a recaptured memory" (Freud, 1937/1976, p. 266).

Unanswered Questions

I have drawn a sharp line between the adaptive unconscious and people's conscious stories about themselves. But the distinction is probably not as sharp as I have implied. People's self stories are undoubtedly a complex mixture of conscious inferences and implicit assumptions that are not easy to verbalize. Consider first-year college students who unexpectedly receive a bad grade in a course. The way that they explain this to themselves is likely to be crucial to what happens next. If they construct a story that they are hopeless failures who will never succeed in college, they will not fare as well as if they infer that they simply need to try harder.

Many studies have assumed that people's self-stories are crucial in situations such as this, and that relatively minimal interventions can succeed in getting people to redirect their story in a more positive direction (Cohen, Garcia, Purdie-Vaughns, Apfel, & Brzustoski, 2009; Walton & Cohen, 2011; Bugental *et al.*, 2002; Wilson, Shelton, & Damiani, 2002). And many of these studies have had spectacular success in doing so, helping to improve academic achievement, reduce child abuse, and alleviate stereotype threat (Wilson, 2012). But, evidence for the proposed mediating processes – namely that people's stories changes in the ways that the researchers assumed – has been hard to come by. Some studies find evidence for story change (e.g., Walton & Cohen, 2011), whereas others do not (e.g., Wilson & Linville, 1982). Of course, this could be because the researchers are wrong about what is driving the change in people's behavior. I suspect that they are right, but that people's stories are a complex mixture of conscious and unconscious assumptions that are not easy to measure.

Second, the pendulum may have swung too far toward the study of unconscious processing. Consciousness has gotten a bad name in some quarters, as researchers (including myself) have investigated its limits and the

problems that result from too much introspection (e.g., Wilson, Dunn, Kraft, & Lisle, 1989). There is a rich tradition of studying the contents of consciousness, dating back at least to James (1890/1950), followed by such endeavors as the study of daydreams (Klinger, 1990; Singer, 1975) and mind wandering (Smallwood & Schooler, 2006). Many of these research programs examine important questions of mental control, or the extent to which people can consciously direct their thoughts, feelings, and behaviors (Wegner & Pennebaker, 1993). Again, much of this work involves the complex interplay of conscious and unconscious processes, such as the extent to which people can keep their mind on a task, and the conditions under which their attention wanders involuntarily from one topic to another. Clearly the interweaving of conscious and unconscious processes is a rich topic to study, though one that requires clever experimental methods.

Summary

Self-knowledge has been a central topic of study for philosophers, religious scholars, and psychologists, as well as for all human beings who have paused for a moment and directed their attention inward. Modern psychology has made considerable strides in understanding the limits of introspection, how self-knowledge can be obtained indirectly through the development of self-stories, and the value of self-knowledge. There is much to be learned about the complex interplay of conscious and unconscious mental processes, and psychological scientists are uniquely equipped to advance this learning.

References

Augustinian Spirituality (n.d.). Prayers of St. Augustine. Retrieved Dec. 4, 2012 from www1.villanova.edu/villanova/mission/campusministry/spirituality/resources/spirituality/restlesshearts/prayers.html

Bugental, D.B., Ellerson, P.C., Lin, E.K., Rainey, B., Kokotovic, A., & O'Hara, N. (2002). A cognitive approach to child abuse prevention. *Journal of Family Psychology, 16*, 243-258.

Cohen, G.L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the achievement gap. *Science, 324*, 400-403.

Colvin, C.R., Block, J., & Funder, D.C. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology, 68*, 1152-1162.

Cooley, C.H. (1902). *Human nature and the social order.* New York: Charles Scribner's Sons.

Dufner, M. (2012). *A differentiated analysis of the social consequences of self-enhancement*. Doctoral dissertation, Humboldt-Universität zu Berlin.

Dunning, D., Griffin, D.W., Milojkovic, J.D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Per-*

sonality and Social Psychology, 58, 568-581.
Flanagan, O. (2011). *The Bodhisattva's Brain: Buddhism naturalized*. Cambridge, MA: MIT Press.
Freud, S. (1976). Constructions in analysis. In J. Strachey (Ed.), *The complete psychological works* (Vol. 23). New York: Norton (originally published 1937).
Gergen, K.J., & Kaye, J. (1992). Beyond narrative in the negotiation of therapeutic meaning. In S. McNamee & K.J. Gergen (Eds.), *Therapy as social construction. Inquiries in social construction* (pp. 166-185). London, England: Sage.
James, W. (1950). *The principles of psychology*. New York: Dover (originally published 1890).
Jones, E.E. (1979). The rocky road from acts to dispositions. *American Psychologist, 34*, 107-117.
Klinger, E. (1990). *Daydreaming*. Los Angeles: Tarcher.
Kross, E. (2009). When the self becomes other: Toward an integrative understanding of the processes distinguishing adaptive self-reflection from rumination. *New York Academy of Sciences, 1167*, 35-40.
Mead, G.H. (1934). *Mind, self, and society*. Chicago: University of Chicago Press.
Pennebaker, J.W. (2004). *Writing to heal: A guided journal for recovering from trauma & emotional upheaval*. Oakland, CA: New Harbinger Publications.
Pronin, E., Gilovich, T.D., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review, 111*, 781-799.
Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 173-220). Orlando, FL: Academic Press.
Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In T. Brown, E.S. Reed & E. Turiel (Eds.), *Values and knowledge* (pp. 103-135). Hillsdale, NJ: Erlbaum.
Schriber, R.A., & Robins, R.W. (2012). Self-knowledge: An individual-differences perspective. In S. Vazire & T.D. Wilson (Eds.), *Handbook of self-knowledge* (pp. 105-127). New York: Guilford.
Schultheiss, O.C., & Strasser, A. (2012). Referential processing and competence as determinants of congruence between implicit and explicit motives. In S. Vazire & T.D. Wilson (Eds.), *Handbook of self-knowledge* (pp. 39 -62). New York: Guilford.
Singer, J.L. (1975). *The inner world of daydreaming*. New York: Harper & Row.
Smallwood, J., & Schooler, J.W. (2006). The restless mind. *Psychological Bulletin, 132*, 946-958.
Solomon, S., Greenberg, J., & Pyszczynski, T. (2004). The cultural animal: Twenty years of terror management theory and research. In J. Greenberg, S.L. Koole, & T. Pyszczynski (Eds.), *Handbook of Experimental Existential Psychology* (pp. 13-34). New York: Guilford Press.
Spence, D. P. (1982). *Narrative truth and historical truth*. New York: Norton.
Saint Teresa of Avila (1577). *The interior castle*. (Translated by Thomas Baker, 1921). Christian Classics Library. Retrieved December 6, 2012 from: http://www.ccel.org/ccel/teresa/castle2.i.html
Taylor, S.E. & Brown, J.D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193-210.
Vazire, S., & Carlson, E.N. (2010). Self-knowledge of personality: Do people know themselves? *Social and Personality Psychology Compass*, 4, 605-620.
Vazire, S., & Wilson, T.D. (Eds.) (2012). *Handbook of self-knowledge*. New York: Guilford.
Walton, G.M. & Cohen, G.L. (2011). A brief social-belonging intervention improves academic and health outcomes of

minority students. *Science, 331,* 1447-1451.

Wegner, D.M., & Pennebaker, J.W. (Eds.) (1993). *Handbook of mental control.* Englewood Cliffs, NJ: Prentice-Hall.

Wilson, T.D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Harvard University Press.

Wilson, T.D., Dunn, D.S., Kraft, D., & Lisle, D.J. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 22, pp. 287-343). Orlando, FL: Academic Press.

Wilson, T.D., & Gilbert, D.T. (2008). Explaining away: A model of affective adaptation. *Perspectives on Psychological Science, 3*, 370-386.

Wilson, T.D., Damiani, M., & Shelton, N. (2002). Improving the academic performance of college students with brief attributional interventions. In J. Aronson (Eds.), *Improving academic achievement: Impact of psychological factors on education* (pp. 88-108). San Diego, CA: Academic Press.

Wilson, T.D., & Linville, P.W. (1982). Improving the academic performance of college freshmen: Attribution therapy revisited. *Journal of Personality and Social Psychology, 42*, 367-376.

Seven Ways Neuroscience Aids Law

Owen D. Jones*

The Game

Law is stuffy, bookish, and boring. Or so many people think. But forget, for a moment, the impressions of law that often come first to mind. Wood-paneled courtrooms. Dusty texts. Numbingly impenetrable language. Gesticulations at a podium.

Now think of law instead as a thrilling, massively multi-player game in which vast resources are at stake, alliances form and dissolve, betrayals and cheating are ever-present threats, arsenals are verbal as well as physical, and the safety of self and property must be navigated with care, en route to some shifting and precarious approximation of peace, happiness, and the acquisition of cool homes, reputations, and gadgets.

It is not my purpose, in prompting this mental shift, to trivialize or glamorize. It is to help us view law from 1,000 meters, so as to see the larger themes – and action-packed patterns – that reveal and define the deeper realities.

In a meaningful way, law *is* a game. There are rules. They are numerous. They are complicated. People have goals that cannot all be satisfied. Choices must be made, turns taken, and consequences – both predictable and not – endured. Each move inspires countermoves that, as in chess, change the strategic landscape of the game forever. And each player adapts to both minutely local and broadly systemic developments. The ecology of law is therefore and inevitably dynamic, fluid, high-stakes, and (often) intensely strategic. And far from being mere entertainment, the game is deadly serious.

This fundamental drama, this surge of human cooperation and competition, is not confined to the familiar domains of crime-scenes and courtrooms. It also pulses out along the web-like threads, in tangles too thick to trace, that constitute the multiple social affiliations of friends, colleagues, businesses, religions, interest-groups, political parties, and the like that so distinctly characterize our ultra-social species.

* New York Alumni Chancellor's Professor of Law & Professor of Biological Sciences, Vanderbilt University; Director, MacArthur Foundation Research Network on Law and Neuroscience.

Law is of course not the only source of potential order in this elsewise chaos, a circumstance vividly and physically illustrated by and within the Vatican City walls that circumscribe our conversation here today. Moreover, there can be wisdom in crowds. For even outside, before, or beyond legal prescriptions, people try things. They notice common interests. And various norms ensue.

But law is *a* source of order, whether it serves a regularizing function (everyone must drive on the same side of the road), an exchange-facilitating function (know ye this: here are the enumerated prerequisites for a contract to be enforceable), a peace-securing function (don't hit people), or the like. And although law is of course a form *of* human behavior it is also manifestly – and most importantly – a system *for regulating* human behavior.

Which brings us to brains.

The Brains

The human animal is numbered in multiple billions. And the game needs certain minimum efforts at order to keep physical force or possession from being the only human measures that matter. Law exists because there are significant interstitial spaces in the behavioral landscape between those places where everyone is already behaving just fine. More pointedly: Law exists mainly to effect *a change in behavior* from how people would have been behaving in the absence of legal intervention – that is, in the absence of some message to the public either about the content of some legal policy (such as: now you may deduct from taxable income the value of your charitable contributions) or about the tools law will use to incentivize behavioral change (such as: from now on, all fines for speeding are tripled).

Which is where brains – and ultimately neuroscience – come in. Because law is, at base, about changing behavior, and because behavior, at base, comes from brains, it follows that deeper understandings of the relationships between brains and behaviors (and, relatedly, about perception, judgment, decision-making, and the like) may aid efforts to increase the effectiveness, efficiency, and justness of law (Jones & Goldsmith, 2005).

To illustrate this central point, I have in the past invoked the metaphor of law as a lever (rather than carrot or stick) for moving behavior. That lever has, as its mission-critical fulcrum, a behavioral model (Jones, 1997; Jones & Goldsmith, 2005). The behavioral model, in turn, is comprised of the algorithms (formal or, much more commonly, informal) by which we predict that if law moves *this* way people are most likely to respond in *that* way, rather than in some *other* way. The key point is that when the lever of law is paired with an inaccurate behavioral model it makes for a soft and inefficient fulcrum at best, and it unintendedly yields disastrous behaviors at worst.

This image helps to explain why the legal system could be described (in a different metaphor) as a monstrously large and ravenously hungry consumer of the various behavioral models offered up by other disciplines. In law theory meets practice. And therefore lives – as well as property, businesses, and even regimes – are lost, saved, or changed when behavioral models are wrong, weak, or insufficient given the leverage society needs.

Law's behavioral models, historically and indeed to date, obviously vary a bit across the specific behaviors at issue, and sometimes reflect multiple models simultaneously. But, broadly speaking, they have included such things as: folk-psychological models (in which a person's actions can be predicted as a function of his or her desires and beliefs) (Nichols, 2002; Davies & Stone, 1995; Morse, 2008, 2011a); informal projections, from introspection, about other people's emotional, analytic, and motivational realities (e.g., "What would incentivize me?"); the sequential installation of one dominant disciplinary view (such as economics) after another (typically to the exclusion of other disciplinary perspectives) (Jones & Goldsmith, 2005); or an amalgam of several close cognate disciplines (such as psychology and sociology) (Jones & Goldsmith, 2005).

But we haven't yet learned nearly enough about why humans behave as they do and what can inspire them to change – as each day's news, whether international or local, so depressingly details. My view has long been that improving the behavioral models on which law relies requires far more aggressive efforts than we've deployed so far to formulate workable syntheses of balkanized fields, not only across the divide that separates the sciences and the humanities, but also within each of those too-separate domains (Jones & Goldsmith, 2005).

I am not the first to observe that the pace of discovery and the volume of ever-increasing knowledge have driven individuals into knowing more and more about less and less, as depth of knowledge inevitably trades against breadth (e.g., Wilson, 1998). Within biology, for example, those studying the same animal at either the ethological or the molecular level can barely talk to one another anymore, so great now are the divides between the methods, vocabularies, and knowledge of these subfields. But for models of human behavior to be accurate and robust, some inward-pulling force must bring the human behavioral disciplines to the very same table. And to that (round) table, I am convinced, must be more systematically added the various life science disciplines – to join with the existing social science disciplines on which law has to date too-exclusively depended.

Species-typical brains develop, through the construction and activity of neurons, at the intersection of genes and environments, as those encounters

are shaped by evolutionary processes over evolutionary time. Behavior is fundamentally a biological phenomenon. Consequently, developing a much deeper and more interdisciplinarily coherent understanding of why people behave in the patterns they do may aid our collective pursuit of an increasingly fair, effective, and just legal system. That may (and in my view probably does) require, at a minimum, increased attention to such fields within behavioral biology as behavioral genetics, evolutionary biology, evolutionary psychology, and neuroscience. Our purpose here focuses on the latter, to which I turn next.

The Relevance of Neuroscience to Law

There are two primary ways that neuroscience can be relevant to law: 1) it can pose new problems; and 2) it can offer aid in solving existing problems.

New Problems: Neuroscience and Legal Headaches

Neuroscience can pose new problems in the same way that many so-called "disruptive" technologies can. For example, given that technologies (such as functional magnetic resonance imaging (fMRI)) enable us to learn something about a person's brain *non-invasively*, could a brain scan by a government ever be characterized as an illegal search and seizure (Farahany, 2012; Shen 2013)? Is a brain scan more like taking a photograph of a person, or more like taking a hair sample for DNA analysis purposes?

When drugs that can enhance cognitive abilities (such as modafinil, which can be used medically for sleep disorders or attention deficit hyperactivity disorder (ADHD)) are used in non-medical settings, by people without medical need, to improve performance on competitive tasks (which is reportedly happening with increasing frequency) should the legal system sit idly by, or instead develop new responses?

What about non-drug technology that can enhance performance, such as the so-called implantable "neuroprosthetics" under development (Hampson *et al.*, 2012; Donoghue *et al.*, 2007)? To what extent should law – distinct from other forms of social order (such as religion, norms, or the like) grapple with the implications?

Or suppose the victim of a violent crime immediately takes a memory-dampening drug, such as propranolol. Should the legal system discount the victim's subsequent identification of a suspect? In an analogous civil case, should the system discount a plaintiff's amount of pain and suffering, on the theory that the drug diminishes it? Should those who failed to take the drug have their recoveries discounted, for failure to mitigate their damages (Kolber, 2006)?

New technologies can also pose distinct problems for judges who must divide the admissible evidence from the inadmissible. In the United States, for example, the last several years has seen both state and federal judges grappling with whether to admit into evidence the results of lie-detection tests that used fMRI technology (*US v. Semrau*; *Maryland v. Smith*).

So one key thing to watch for, in coming years, is instances when law is forced to face novel questions driven by novel technologies. For now, however, I want to focus our attention on the second kind of relevance: when neuroscientific insights or findings may provide direct value to the legal system, as it goes about trying to pursue the goals society assigns to it.

New Aid: Seven Values to Law of Neuroscience

As I see it, neuroscience can provide value to law in at least seven different ways.

Category 1: Buttressing

The scientific enterprise is one that – in the broader scheme of things – takes as long as it takes to get things as right as it can. Experiments provide the evidence on which best approximations are based. And inconclusive experiments can be followed by different or better experiments.

In contrast, the legal system rarely operates within an explicitly experimental paradigm. Instead of going back to the drawing board until more knowledge can be acquired, policy-makers, judges, and others in the legal system must often and quite unavoidably take definite action when the state of knowledge remains something far, far short of the scientifically familiar $p \leq 0.05$.

A defendant must be convicted or let go. A plaintiff who may have been injured by a corporation's action must succeed and get paid or not. New investigations for finding further facts would be just lovely to have. But the facts are generally either not amenable to discovery through experimental means (as in: was the driver negligent in not noticing the pedestrian crossing the road in the night?) or, even when new facts are amenable to discovery, courts are simply not in the business of designing or ordering new experiments in furtherance of truth.

In law, in addition, the best decision that can be made, with the limited evidence one has, must be made on the timeline society expects. Trials are not designed to discover new truths about the world, as science is. As David Faigman has succinctly put it: "While science attempts to discover the universals hiding among the particulars, trial courts attempt to discover the particulars hiding among the universals" (Faigman, 1999). Trials are therefore

designed to provide the most just result that can be reached, regarding a typically unique set of facts … *given* the constraints of time, resources, uncertainties, and the vagaries of evidence – which is generally and regrettably dependent on the self-reported memories of parties who are imperfect at best, and self-interested at worst.

The consequence of all this is that decisions are often made on the basis of evidence that leans *sufficiently* strongly in one direction. (In the United States, for example, different decision thresholds – such as *beyond a reasonable doubt* or *by a preponderance of the evidence* – formally accompany different legal contexts). But because many different forms of evidence can be weighed (and often are weighed) together, neuroscientific evidence (assuming the judge finds it relevant and not unduly prejudicial to unbiased decision-making) can fit quite comfortably within the system, just like any other form of evidence to consider and weigh. My point is that neuroscientific evidence – whether advanced in individual trials, for legislative purposes, or the like – is rarely if ever going to be the sole form of evidence that is relevant to a legal decision.

In some cases, for example, neuroscience might "only" point in the same direction as other evidence. And in those cases we might call this a "buttressing" function, inasmuch as the neuroscientific evidence collaterally supports, and further strengthens, something that already stands independently (or nearly so).

But even if the neuroscientific evidence doesn't change the *outcome* in such a circumstance, it is worth noting that it still changes the context. And that can provide an important advantage to the legal system – in the same way that four different and independent methods for reaching a similar conclusion can provide better support for that conclusion than would just one or two methods alone. (Recall, for instance, the famous triangulation from multiple research streams that Charles Darwin deployed in his *Origin of Species* (1859)).

Here is a concrete example. Suppose a person behaved in a criminally violent fashion, was arrested, and is now to be tried. The circumstances are bizarre and seemingly motiveless. The defendant looks "off" – in the eyes and facial expressions – even to strangers viewing still photographs of him. Witnesses attest that the defendant not only behaved out of the norm for law-abiding people, but also out of the norm for those who are criminally violent. Other witnesses report that the defendant, for at least the past year, was in such a worsening state of bizarre behavior that mental illness was strongly suspected. A clinical psychiatrist interviewing the defendant concludes that he is medically insane. And – to the point – a structural brain

scan (such as an MRI), a functional brain scan (such as PET), or both, reveal an extremely large tumor impinging on those regions of the prefrontal cortex commonly associated with the ability to inhibit and self-regulate.

In such a circumstance, what can we say of the value to law of the neuroscientific evidence? What we can say is that it can be quite valuable, even if it is not what lawyers call "dispositive" by itself. That is, by triangulating with other forms of evidence, the neuroscience increases confidence in the conclusion that the defendant was less responsible for his (not necessarily excusable) behavior than is the average similarly-aged citizen. Although this does not tell us what to do with such a person, the neuroscientific buttressing adds value to our deliberations.

Category 2: Challenging

In contrast, there may be times when neuroscientific evidence is either in tension with some other evidence or is in tension with some assumptions undergirding the entire legal context and approach. This function of neuroscience – which could run along a continuum from *calling-into-question* to *challenging* to outright *contradicting* – can add value to law's efforts to avoid error. That is, the neuroscientific findings may prompt useful course-corrections in a legal procedure, or may even prompt reforms in a legal policy or approach (Morse, 2011b).

Suppose, for example, that an assumption that underlies a particular feature of law is incorrect. A colleague of mine, Michael Vandenbergh, noticed that a provision of the *Federal Rules of Evidence* in the United States provides a valuable heuristic in this regard. Those rules prohibit, generally speaking, the admission of Bob's testimony – regarding a statement made by Charles about the existence of Fact X – if Bob's testimony is being introduced in an effort to prove that Fact X exists. (Fed. R. Evid. 801(c)). The principle rationale for the rule excluding such testimony is that the opposing party can't confront Charles directly about the basis of his own statement, which is merely "hearsay" when offered by Bob.

Yet among the exceptions to the exclusion of hearsay evidence is one for so-called "excited utterances". An excited utterance is "A statement relating to a startling event or condition, made while the declarant was under the stress of excitement that it caused" (Fed. R. Evid. 803(2)). Excited utterances are admissible – even when offered by those who merely overheard them – because they are considered unlikely to be deceptive. As the U.S. Supreme Court stated in *Idaho v. Wright*, "[t]he basis for the 'excited utterance' exception ... is that such statements are given under circumstances that eliminate the possibility of fabrication, coaching or confabulation, and

that therefore the circumstances surrounding the making of the statement provide sufficient assurance that the statement is trustworthy and that cross-examination would be superfluous".

The excited utterance example highlights the legal system's frequent dependence – typically without knowing it – on what are, at base, *neuroscientific* assumptions. Beneath the excited utterances exception to the hearsay rule, for example, is the implicit assumption that human brains just don't work with sufficient speed, or at least with sufficient speed when one is in a state of excitement, to afford a reasonable opportunity to lie. Leaving aside the fact that being in a state of excitement may very well interfere with the accuracy of one's perception, the legally operative assumption that people can't lie quickly when excited is testable (at least in theory). And it may very well be completely wrong. Consequently, any strong neuroscientific showing to that effect could potentially lead to valuable reform of the legal rule.

A wide variety of other assumptions in the evidentiary rules could similarly be called into question, given appropriate neuroscientific findings. For if there is one place in the legal system where rules are quite systematically a function of our shared understandings of what will probably happen in another person's brain it is in the evidentiary rules. Those rules enable a judge to prevent jurors from seeing or hearing certain kinds of information when there is reason to believe that jurors are likely to have their judgments compromised in one or more ways. Put another way, the evidentiary rules are quite inevitably built on assumptions about what happens, and with what consequences, when certain kinds of information get into a juror's brain. (For more, see Brown, 2012).

This example, within the context of evidentiary rules, is but one illustration of the broader point I want to make about the *Challenging* category. And that is that any time well-executed and properly interpreted neuroscientific developments strongly challenge important assumptions on which a given feature of law relies it can add value by virtue of such challenges. It may help the legal system avoid error, and might prompt useful reform of approaches based on faulty assumptions.

Category 3: Detecting

Into my third category fall all manner of valuable contributions of neuroscience that involve detecting, better than could previously be done, facts that are legally relevant. Better methods for detecting the extent of brain injuries (through structural and/or functional scans) would be a paradigmatic example. So, too, would be the use of fMRI or EEG for lie detection purposes (should the techniques advance that far) (Wagner, 2010; Shen &

Jones, 2011; Greely & Illes, 2007). The same is true for detecting autobiographical memories (say, recognition of pictures of a particular terrorist), or for detecting and quantifying subjective pain.

Those illustrations (repeated within this larger set, for convenience) multiply easily:

- How injured is this person's brain, and with what functional consequences?
- Is this person lying?
- Does this person recognize this target stimulus (e.g., a person's photograph)?
- How much pain and distress is the litigant feeling?
- What capacity did this person have to act differently than she did?
- Is this person mentally ill?
- What does this person remember?
- How accurate is this person's memory?
- What was this person's probable mental state, at the time of the act?

Adding to the corpus of potentially relevant facts, by detecting things that are otherwise undetectable, can in suitable instances aid the legal system's efforts to answer some of the big and perennial questions.

Category 4: Sorting

The legal system is frequently called upon – in both criminal and civil contexts – to sort people into categories, such as free or incarcerated, death-penalty-eligible or not, sane or insane, deserving of compensation or out of luck, … even dead or alive. In some contexts, neuroscience can offer aid with that sorting.

In criminal contexts, for example, one of the hardest problems to solve in law is this: How do you minimize the sum of the total societal costs of incarcerating a criminal and later returning him to society? Longer incarceration is costly (in taxes); but so is recidivism (in criminal activity). The latter costs are more vivid to the public, though not necessarily greater. One of the ways neuroscience can help the legal system minimize the total costs is by improving law's ability to sort people into groups that perhaps should be treated differently under the law. For example, neuroscience might help to illuminate the extent to which a particular defendant, or indeed certain kinds of defendants (addicts, for example, or juveniles), might respond more effectively to medical treatment, or to special treatment, than to standard-style incarceration.

In civil contexts, neuroscience can provide, for example, and indeed has provided, one useful measure, "brain death", for sorting the legally alive

from the legally dead (e.g., New Jersey Declaration of Death Act). Given the shortage of organs available for donation, and the advantages of removing organs from a body that is still metabolic active, defining the precise moment of legal death has very important legal (as well as social and medical) ramifications.

Category 5: Intervening

Another way neuroscience can help the legal system minimize the total costs of incarceration and reintegration is by offering new and effective interventions. For example, psychopharmacological neuroscience might present new drugs, with new capabilities, that can help to meaningfully reduce the incidence of certain kinds of recidivism.

Interventions come in many kinds. And the use of neuroscience to recommend an intervention does not necessarily imply that the intervention must itself also be neuroscientific – as would be psychopharmacological intervention or even (in theory) invasive techniques similar to those used in the deep brain stimulation (DBS) techniques for combatting Parkinson's disease or major depression. Many valuable interventions will continue to be strictly behavioral ones, for example through targeted courses of education and training, or through specialized techniques of behavior modification (also known as applied behavior analysis (ABA) or positive behavior support (PBS)).

Category 6: Explaining

There are times when neuroscience may help to explain or illuminate matters that, though not actively contested, have previously been beyond the reach of technological investigation. For example, colleagues and I at Vanderbilt University have used neuroscientific methods (fMRI) to reveal the brain activity underlying the decisions whether or not to punish and, if so, how much (Buckholtz *et al.*, 2009; Jones *et al.*, 2009). Studies like these, which may illuminate brain regions and neural patterns correlated with various aspects of legal decision-making, may ultimately help the legal system to learn more about pathways by which inappropriate biases can infect decision-making. And this in turn might facilitate efforts to combat such biases. Similar advances in explaining other law-relevant phenomena may provide the basis for adding value in other categories as well, such as in pointing a path toward effective interventions, or toward improved prediction of future behavior.

Category 7: Predicting

Neuroscience can also help law to the extent it can improve law's ability to make predictions. Some of these predictions may be about the amenability of a particular defendant to treatment (demonstrating, by the way, the inherent overlap between categories – in this case between *Predicting* and *Sorting*), or about the future behavior of individuals or groups.

For example, some of the most challenging questions the legal system routinely faces are ones like this: Given what this person did, the circumstances under which he did it, his nature as nearly as can be discerned, and his behavior while in custody, what is the likelihood that, if released, he will commit another violent crime? Biomarkers (structural, functional, or both) that neurosciences may discover might – in combination with other forms of knowledge, such as behavioral genetics, social psychology, clinical psychiatry, and the like – improve the accuracy of predictions about recidivism (e.g., Aharoni, *et al.*, 2013). In such a case, neuroscientific information could add considerable value to parole decisions, for instance.

The Non-Delegable Duty

This very brief taxonomy of some of the ways that neuroscience may prove valuable to law is subject to one over-arching principle. And that is that the legal system cannot delegate to another field, scientific or otherwise, the ascription of legal meaning. By this I mean that – even assuming that scientific testimony is unanimous as to a particular fact – the *legal* meaning of that fact is inevitably, unavoidably, and unshirkably a decision that legal decision-makers must bear. This is, in some respects, a formal legal application of *The Naturalistic Fallacy* (Hume, 1978; Moore, 1993), which underscores the important point of logic that facts never speak, all by themselves, to the appropriate normative conclusions. Put another way: an "is" cannot, by itself, yield an "ought", for the same reason that explanation is not justification, and that description is not prescription.

This is not to exaggerate the importance of law. I see neither law nor science as superior to the other, inasmuch as they have very different domains, purposes, and tasks. But it is important to draw a distinction between neuroscience answering a legal question and neuroscience *helping* to answer a legal question. At the same time that I think the former will be quite rare, I think opportunities for the latter are considerable, and quickly increasing.

The Endgame

The endgame of this discussion has one clarification and four main points. The clarification concerns the scope and nature of the taxonomy. Specifically, these seven usages – though distinct from one another – can and often will overlap in specific instances, given that a single neuroscientific result can often be used for more than one purpose.

Put another way, this set describes seven *forms of relevance*; it should not be thought to provide a rigid, over-reifying set of mutually exclusive cells in some hypothetical matrix that would divide the entire logical space (the various portions of which obviously can, in any event, be organized in multiple useful ways, see, e.g. Morse, 2011b). My four main points are these.

First, law needs neuroscience for the same reason it needs other life sciences: to deepen its understanding of the human animal in ways that may lead to more effective, efficient, and just regulation of human activities. A large and growing literature is making important contributions to this domain (*Law & Neuroscience Bibliography* (2013); Shen 2010; Morse & Roskies (in press)).

Second, the legal system will never reach its maximum potential if the behavioral models on which it relies are less complete than they might be, given further efforts to force reconciliation of differing disciplinary perspectives. Those who work in law should do all they can to encourage improved syntheses of disciplinary perspectives in the arenas of behavior most directly relevant to law.

Third, the legal system needs to gear up to confront the new legal problems and questions that neuroscience increasingly offers (Jones, Schall, and Shen, 2014). These questions are already appearing in the courts, where ignoring them is not an option (Jones & Shen, 2012). And this flow of new questions raised by neuroscientific techniques is far more likely to increase than to decrease.

Fourth and finally, there are a number of distinct ways – including at a minimum in the contexts of *Buttressing, Challenging, Detecting, Sorting, Intervening, Explaining, and Predicting* – that neuroscience can offer value to law. Though each of these contexts can afford real and tangible advantage to the fair and effective administration of justice, the specific pathways in which it may do so – and with what specific downstream implications – are yet to be fully identified and navigated.

Acknowledgments

The preparation of this paper, and of the November 2012 talk at Vatican City on which it is based, was aided in part by a grant from the John D. and Catherine T. MacArthur Foundation, as well as from discussions with Jeffrey

Schall, Francis Shen, Gideon Yaffe, co-participants in the Working Group on "Neurosciences and the Human Person" of the Pontifical Academy of Sciences, and numerous colleagues within the *MacArthur Foundation Research Network on Law and Neuroscience* (www.lawneuro.org). Sarah Grove and Sonal Patel provided able research assistance. Correspondence should be sent to owen.jones@vanderbilt.edu.

References

Aharoni, E., Vincent, G.M., Harenski, C.L., Calhoun, V.D., Sinnott-Armstrong, W., Gazzaniga, M.S., & Kiehl, K.A. (2013). Neuroprediction of future rearrest. *Proceedings of the National Academy of Science, 110*, 6223.

Brown, T. (2012). The psychological presumptions underlying the federal rules of evidence. (working paper).

Buckholtz, J., Jones, O., Asplund, C., Dux, P., Zald, D., Gore, J., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron, 60*, 930.

Darwin, C. (1859). *On the origin of species by means of natural selection*. London: John Murray, Street.

Davies, M. & Stone, T. (eds). (1995). *Folk psychology: the theory of mind debate*. Cambridge, MA: Blackwell.

Donoghue, J.P., Nurmikko, A., Black, M. & Hochberg, L.R. (2007). Assistive technology and robotic control using motor cortex ensemble-based neural interface systems in humans with tetraplegia. *Journal of Physiology*. Special Issue on Brain Computer Interfaces*, 579,* 603-11.

Faigman, D. (1999). *Legal alchemy: the use & misuse of science in the law*. W.H. Freeman.

Farahany, N. (2012). Searching secrets. *University of Pennsylvania Law Review, 160*, 1239.

Greely, H.T. & Illes, J. (2007). Neuroscience-based lie detection: the urgent need for regulation. *American Journal of Law & Medicine, 33*, 377-431.

Hampson, R., Gerhardt, G., Marmarelis, V., Song, D., Opris, I., Santos, L., Berger, T., & Deadwyler, S. (2012). Facilitation and restoration of cognitive function in primate prefrontal cortex by a neuroprosthesis that utilizes minicolumn-specific neural firing. *Journal of Neural Engineering, 9*, 1-17.

Hume, D. (1978). *A treatise of human nature*. Dover Publications: Mineola, New York. *Idaho v. Wright*, 110 S. Ct. 3139, 3149 (1990).

Jones, O. (1997). Law and biology: toward an integrated model of human behavior. *Journal of Contemporary Legal Issues 8,* 167-208.

Jones, O. & Goldsmith, T. (2005). Law and behavioral biology. *Columbia Law Review, 105*, 405.

Jones, O., Buckholtz, J., Schall, J., & Marois, R. (2009). Brain imaging for legal thinkers: a guide for the perplexed. *Stanford Technology Law Review, 5*.

Jones, O. (2004). Law, evolution and the brain: applications and open questions. *Philosophical Transactions of the Royal Society, 359*, 1697-1707.

Jones, O., Schall, J., & Shen, F. (forthcoming 2014). *Law and neuroscience*. Aspen Publishers.

Jones, O. & Shen, F. (2011). Law and neuroscience in the United States. In Tade Spranger (Ed.), *International neurolaw: a comparative analysis* (pp. 349) Springer-Verlag.

Kolber, A. (2006) Therapeutic forgetting: the legal and ethical implications of mem-

ory dampening. *Vanderbilt Law Review*, *59*, 1561.

Law & Neuroscience Bibliography, of the *MacArthur Foundation Research Network on Law and Neuroscience*, available at: www.lawneuro.org/bibliography.php

Mobbs, D., Jones, O., Lau, H.C., & Frith, C.D. (2007). Law, responsibility, and the brain. *PLOS: Biology, 5*, 693.

Moore, G.E. (1993). *Principia ethica*. Cambridge University Press.

Morse, S. (2011a). Avoiding irrational neurolaw exuberance: a plea for neuromodesty. *Mercer Law Review*, *62*, 837.

Morse, S. (2011b). The status of neurolaw: a plea for current modesty and future cautious optimism. *Journal of Psychiatry & Law, 39*, 595.

Morse, S. (2008). Determinism and the death of folk psychology: two challenges to responsibility from neuroscience. *Minnesota Journal of Law, Science & Technology, 9*, 1-36.

Morse, S. & Roskies, A. (eds) (in press 2013). *A Primer in Criminal Law and Neuroscience*. Oxford.

New Jersey Declaration of Death Act, N.J.S.A. 26:6A-1.

Nichols, S. (2002). Folk psychology. In *Encyclopedia of Cognitive Science*. Nature Publishing Group, 134-140.

Shen, F. (2013). Neuroscience, mental privacy, and the law. *Harvard Journal of Law & Public Policy*, *36*, 653.

Shen, F. (2010). Law and neuroscience bibliography: comments on an emerging field. *International Journal of Legal Information, 38*, 352.

Shen, F. & Jones, O. (2011). Brain scans as evidence: truths, proofs, lies, and lessons. *Mercer Law Review 62*, 861.

State of Maryland v. Gary Smith, Sixth Judicial Circuit, No. 106589C (2012).

U.S. v. Semrau, U.S. District Court for the Western District of Tennessee. No. 2:07-CR-10074-JPM (2011).

U.S. v. Semrau, U.S. Court of Appeals for the Sixth Circuit, No. 11-5396 (2012).

Wagner, A. (2010). Can neuroscience identify lies? In Wagner, A. (Ed.), *A judge's guide to neuroscience 13* (pp. 13-25).

Wilson, E.O. (1998). *Consilience: the unity of knowledge*. Knopf.

Interaction Between Two Readings: The Naturalistic and the Socratic "Know Thyself"

Marcelo Sánchez Sorondo

Knowledge about Man: the possibility of two approaches

There was no great problem between the different domains of knowledge until a border was drawn between nature understood as having a soul or surrounded by a soul, and a soul which was in itself characterised by an end: this was the age of Aristotle's *Physics*, *De Anima* and *Ethics*. This border was drawn at the end of the Renaissance, which had not assimilated the originality of the thought of St. Thomas.

The problem became acute when nature became the subject of a science based on pure observation, mathematical calculation, and experimentation. This was the meaning of the Galilean and Newtonian revolution, as Kant (1787) defined it.[1] The human mind thought that it did not have access to the principle of the production of nature in itself or in something other than itself, what Aristotle called form or the formal principle as principle of operation: 'every essence in general is called "nature", because the nature of anything is a kind of essence'.[2] Therefore one can only gather natural gifts made known through their appearance in space and time and try to 'save the phenomena – σώζειν τα φαινόμενα', as Plato himself suggested, who in this was Galileo's mentor. This is no minor endeavour given that the field of observation is so unlimited and that the imaginative ability to form hypotheses with a mathematical formula, to enlarge and replace models, to vary the character of models, and to invent procedures of verification and falsification, is so powerful. This is no minor endeavour, also, because mathematics, which is in part a construction of the mind of the human being, corresponds to the quantity that indeed constitutes the specific matter of every individual and expresses in bodies the realisation of individuality through the parts of such material structure. There is quantity in the mind

[1] Kant, Immanuel, *Critique of Pure Reason*, Preface to the second edition (1787). Available online at http://etext.library.adelaide.edu.au/k/kant/immanuel/k16p/k16p2.html

[2] Aristotle, *Metaph.*, 5, 1015 a 12 f.

of man and in the corporeal structure (atoms and sub-atomic structures, molecules, cells, organs, etc.). Thus, although there is no ancient Aristotelian correspondence between the mind and reality through the notion of form, there is the modern correspondence through quantity inspired in Pythagoras of Samos and Plato – something that has been pointed out on more than one occasion by Benedict XVI in his Magisterium.

However, as regards phenomena relating to human beings, this asceticism of hypotheses, of the creation of models, and of experimentation, is in part compensated for by the fact that we have partial access to the production of certain phenomena that can be observed through philosophical self-reflection (and of course, for believers, through faith). Thus we are dealing with what in the praxes that are different from this scientific theory and technologies can be deemed the genetics of action that belong to fundamental anthropology and to ethics. Reflection on human praxes expresses the point of convergence because it indicates the path that leads to the end, i.e. perfect human work as fullness of the act. The success of work (ἔργόν) can only be observed in the perfection of praxis itself (ἐνέργεια) in relation to its end.

Thus the action shows that every man and every woman as individuals proceeds for an end and thus that she/he himself is the principle of action: '*hic homo singularis intelligit, vult, amat*'. As will be discussed, this is the starting point of St. Thomas against Averroism which anticipates in many ways some decisive modern philosophies and of Ricœur against the "masters of suspicion" (Marx, Nietzsche and Freud) for their methodological and substantial naturalism. In the vast field of activity, the human being considers himself responsible for his own action. This means that he can go back from the observable effects of his actions to the intention that gives them meaning and even to the mental acts which create finalities that generate the intentions and the observable results. Thus the action not only exists to be viewed from the outside, like all the natural phenomena of which it is part: it exists to be understood beginning with expressions that are at one and the same time the effects and signs of the intentions that give meaning to it and with the acts that create meaning that at times sometimes produce such intentions. It follows from this that man's knowledge is not a matter of a single plane or level – that of external observation, explanation, and experimentation (as a reproduction of phenomena) beginning with his body and brain: this knowledge develops in the interface between the observation of nature and reflective understanding. Therefore the human being is simultaneously an observable being, like all the beings of nature in which he participates, and a being who interprets himself, who knows himself as Heraclitus, Socrates, Aristotle, Thomas

Aquinas, Hegel had already suggested (a 'self-interpreting being' to employ the phrase of Charles Taylor or Paul Ricœur).[3]

This statement on the various objective levels of knowledge and of the science of knowledge, or epistemology, and to begin with on the different levels of knowledge and self-awareness of the human being, can provide an answer of reconciliation and pacification to the question raised by the status of the human being in the age of predominance of natural's sciences, as long as, that is, positivist ideology does not claim the right to abolish the border between the sciences of nature and the sciences of man and to annex the latter to the former. Regretfully, contemporary philosophy answered this challenge by simply juxtaposing an abstract anthropology or a phenomenology of the concrete man, without articulating its discourse on the way this acting and suffering being behaves in the world with the scientific discourse. It may be difficult to ask today's philosophers or theologians to become scientists or specialists and vice versa: however, the needs of the condition of contemporary man strongly encourage us to open up to an indispensable participation in interdisciplinary research where theologians, philosophers, thinkers and scientists are willing to work together. We try to do this in our workshop.

The Neurosciences and Self-Understanding

A controversial point to achieve this pacification might be the field of the neurosciences. In terms of this approach, the scientist is expected to seek at the cortical level the correlation between the observable structures and the functions where the structures are the bases, the supports, the nervous material or whatever we may want to call it. The scientist only observes quantitative and qualitative changes, the ever more complex hierarchies of observable phenomena; but the meaning of the function which corresponds to the structure is understood only by the speaking subject who says that he perceives, that he imagines, and that he remembers. These oral statements, together with behavioural signs that the human being shares to a large extent with the higher animals, fall within a type of analysis where there is no mention of neurons, synapses etc. but reference is made to im-

[3] On this point we find an illuminating text in the Encyclical *Fides et ratio* which declares: 'Metaphysics should not be seen as an alternative to anthropology, since it is metaphysics which makes it possible to ground the concept of personal dignity in virtue of their spiritual nature. In a special way, the person constitutes a privileged locus for the encounter with being, and hence with metaphysical enquiry' (§ 83).

pressions, intentions, dispositions, wishes, choices, ideas etc. We again find here a certain semantic dualism, if we can use this phrase, which does not, however, jeopardise the integral knowledge about the human being. An important corollary of such semantic dualism lies in the fact that we speak in similar terms of the body, of the same body, in both analyses: there is the body-object, of which the brain is the guiding force with its marvellous architecture, and the body proper, this body that is the only one that is mine, that belongs to me, which I move, which I suffer; and there are my organs, my eyes 'with' which I see, my hands 'with' which I grasp. And it is on this body proper that all the architecture of my powers and my non-powers is built: the power to do and not to do; the power to do this or that; the power to speak, to act, to attribute to myself my own actions, given that I am their real author, and thus free. In short, I find in my body something radical which is my free capability to act, which in Latin may be defined as *capax*, the human being as capable to act and to be aware and free of it through his body and brain.

There is thus raised the question of the relationship between the two analyses or approaches – that of the neurologist and that of the philosopher and metaphysician. And it is here that the analyses cross over without ever dissolving each other. The scientist and the philosopher can agree on calling the body-object (and its marvel, the brain), the 'reality without which we cannot speak, or think or decide or feel or live or act'. The scientist can continue to have a naturalistic viewpoint in his analysis which enables him to work without direct metaphysical perspectives. The philosopher speaks about the brain in terms of recipient structure, of support, of substrata, of basis, of potency, of encephalic matter, of part of the person. It must be accepted that, for the moment, we do not have a sort of third analysis where there is awareness that this brain-body and my living body are one and the same being. However, the analysis of the brain-body must have a certain opening towards the analysis of my living body and vice versa, namely that while the analysis of my living body gives to me in itself my experience and philosophical reflection, it must be open or enable indirectly or *per accidens* the analysis of the mind-body and vice versa.

We notice here that we do not have direct access to the very origin of the being that we are, in other words we do not have a sort of self-transparency of ourselves and of our selfhood and, starting from this centre, a self-transparency also of all of our actions. In this sense we cannot understand ourselves immediately through our being and essence by essence. On the contrary, our being attests to its existence in the concrete and current exercise of our life. In a realistic clime and vision, St. Thomas indicates this

clearly: 'For one perceives that he has a soul, that he lives, and that he exists, because he perceives that he senses, understands, and carries on other vital activities of this sort' ('*In hoc enim aliquis percepit se animam habere, et vivere et esse, quod percepit se sentire et intelligere et alia huiusmodi opera vitae exercere*').[4] For this reason Aristotle declares: 'We sense that we sense, and we understand that we understand, and because we sense this, we understand that we exist'.[5] In the perception of our praxis or activity there is the co-perception of the beginning: 'from a perception of the acts of the soul we perceive the principle of such acts' ('*perceptis actibus animae, percipitur inesse principium talium actum*')'.[6] St. Thomas assures us that our soul, since it grasps universals, perceives (*percepit*) that is has a spiritual form; he argues that we are aware of the very becoming of the universal in the soul and even that the very light of intelligence makes its presence known to us by means of the soul. This signifies affirming in an explicit manner a perception proper to the spiritual reality in a positive way but by means of the spiritual operation of implementing the intelligible: 'And we know this by experience, since we perceive that we abstract universal forms from their particular conditions, which is to make them actually intelligible' ('*Et hoc experimento cognoscimus, dum percipimus nos abstrahere formas universals a conditionibus particularibus, quod est facere actu intelligibilia*').[7]

The ultimate originality of this perception of our spiritual reality is the absolutely original fundamental situation which we may call the genetics of the act or 'the emergence of freedom' as a move from potency to the act or the capability to free act' or the capability of acting or of non-acting and our awareness of it. Quite rightly Christian thought, long before, and with more precision than, the moderns, when considering this reality of the spiritual subject called freedom the '*motor omnium*' of the activity of the person, and the protagonist of the person, the 'I', the self (selfhood), the human subject that we discover through praxis. This perception is so radical that it is more than an opinion and it is prior to every science, whether theoretical or prac-

[4] St. Thomas Aquinas, *Q. d. De Veritate*, q. 10, a. 8.

[5] Aristotle, *Ethica Nicomachea*, IX, 9, 1170 a 30.

[6] St. Thomas Aquinas, *Q. d. De Veritate*, q. 10, a. 9.

[7] St. Thomas Aquinas, *S. Th.*, I, q. 79, a. 4. Available online at http://www.corpusthomisticum.org/sth1077.html He also states: 'The human soul understands itself through its own act of understanding, which is proper to it, showing perfectly its power and nature' i.e. '*Anima humana intelligit seipsam per suum intelligere, quod est actus proprius eius, perfecte demonstrans virtutem eius et naturam*' (*Ibidem*, I, q. 88, a. 2 ad 3; available online at http://www.corpusthomisticum.org/sth1084.html).

tical; indeed it is converted into the principle of the foundation of the different praxes. We can say that it is a form of belief, a *Glauben*, in the non *doxic* sense of the term, if we reserve the term *doxa* for a degree lower than *episteme* and in the order of the phenomena of nature and also in that of human phenomena liable to being treated they themselves as observable. The belief proper of attestation of our freedom is of another order; it is of the order of conviction and confidence; its opposite is suspicion, not doubt, or doubt as suspicion (P. Ricœur); it cannot be denied, but refused; it cannot be re-established and strengthened if not through resorting again to attestation, and is rescued by the approval of the other, indeed thanks to some kind of gracious divine support. In this context to which fundamental anthropology refers, one can observe that one is dealing with a truth that is closely connected with the fundamental conviction that the human being has of himself and which is not temporary as is the case with the acquisitions of the arts and sciences and philosophy itself with which, however, it has a close relationship, and thus one speaks of 'philosophical anthropology' to refer to its specific genre of knowledge through reflection that takes place by stages.

Brain, Mind, Soul and Being

Aware of the lack of a direct and perfect self-transparent knowledge of such a founding origin, scientists and philosophers should aim to seek an increasingly precise adjustment between a neuroscience which is increasingly expert in material architecture and phenomenological and anthropologic descriptions centred on human operations (seeing, understanding, living well, acting) where praxis is subject to philosophical analysis. So, the point of departure and turning point to both approaches is human praxis. In Aristotle, the act that achieves a human praxis is clearly dissociated form the act of movement (κίνησις) and is, instead, associated in a privileged way with that of action, in the sense of praxis (πρᾶξις): 'Since no action which has a limit is an end, but only a means to the end, as, e.g., the process of thinning; and since the parts of the body themselves, when one is thinning them, are in motion in the sense that they are not already that which it is the object of the motion to make them, this process is not an action, or at least not a complete one, since it is not an end; it is the process which includes the end that is an action. E.g., at the same time we see and have seen, understand and have understood, think and have thought; but we cannot at the same time learn and have learnt, or become healthy and be healthy. We are living well and have lived well, we are happy and have been happy, at the same time; otherwise the process would have had to cease at some time, like the thinning-process; but it has not ceased at the present moment; we both are living

and have lived. Now of these processes we should call the one type motions, and the other actualisations. Every motion is incomplete – the processes of thinning, learning, walking, building – these are motions, and incomplete at that. For it is not the same thing which at the same time is walking and has walked, or is building and has built, or is becoming and has become, or is being moved and has been moved, but two different things; and that which is causing motion is different from that which has caused motion. But the same thing at the same time is seeing and has seen, is thinking and has thought. The latter kind of process, then, is what I mean by actualisation, and the former what I mean by motion'.[8] What makes this text remarkable is that the disjunction between action and movement is upheld by a criterion that involves a phenomenology of a metaphysical character, namely the possibility of saying, 'at the same time', we are seeing and we have seen, we are living well and have lived well, we are happy and we have been happy. The interaction of the tenses of verbs, and in a certain sense its overcoming which is arranged around the difference between movement and human praxis, reveals a fundamental phenomenon that bears upon the temporality of human acting. The fact that the perfect and the present are 'together' implies that everything that the perfect contains of the past is recapitulated in the present and *vice versa*. The human being, therefore, not only is capable of measuring temporal succession according to a first and then but is also, after a certain fashion, above time and the foundation of its succession which is matter. In spiritual operations the human being transcends the movement of nature and lastly matter itself, and is directed, according to the suggestive statement of St. Thomas, towards the Infinite: '*Simpliciter quidem, sicut intelligere, cuius obiectum est verum, et velle, cuius obiectum est bonum, quorum utrumque convertitur cum ente; et ita intelligere et velle, quantum est de se, habent se ad omnia*'.[9] If this kind of praxis transcends pure movement it is because it is a more perfect kind of act, that is to say it has all the perfection of the act of movement but its imperfection is not linked to the succession of matter.[10]

This connects the investigation of the being of the self to the interpretation of one of the four primordial meanings of being, which Aristotle

[8] Aristotle, *Metaph*, IX, 6, 1048 b 18-35.

[9] *S. Th*, Iª, q. 54 a. 2 co.

[10] Cf. Paul Ricœur, 'Tenth Study: What Ontology in View?', in *Oneself as Another* (Chicago-London, 1992), pp. 302-308; 'Que la science s'inscrit dans la culture comme "pratique théorique"', in *The Cultural Values of Science* (The Pontifical Academy of Sciences, Vatican City, 2003), pp. 14-23, available online at http://www.pas.va/content/accademia/en/publications/scriptavaria/culturalvalues.html

placed under the distinction of act and of potency.[11] Now, it is essential for an ontological exploration of human acting, understood as being different from the movement of nature, that the same examples taken from human praxis appear at the same time as *centred* and *decentred*. In others words, if the meanings of being such as *dynamis-energeia* were only another way of saying *praxis,* the metaphysical lesson would be meaningless. And rightly to the extent that *dynamis-energeia* can irrigate other fields of application different from human action that manifests its analogical fecundity. The essential is the decentring itself towards the bottom and towards the top, in Aristotle (and in St. Thomas during the Middle Ages and in Ricœur in our contemporary time), in virtue of which the *dynamis-energeia* indicates a basis of being, at one and the same time powerful and effective, on which human acting stands out. In other terms, it appears equally important that human acting is the privileged place of the readability of this meaning of being because it is distinct from all the acts of physical nature, and that being as act and potency has other fields of actuation that are different from human acting. Centrality of acting and decentring or better *re-centring* in the direction of a basis of act and potency which Aristotle himself defines as 'first act' because it is distinct from all the others, when it is a matter of explaining the soul as form. This analogy attests that for Aristotle the being of man is that basis of being as first act starting with which he can be an agent and receiver that transcends matter and is capable of measuring time.

And it is here, in that higher sphere of human praxis which is knowing, that Aristotle distinguished in a new way for the first time two acts and two potencies: the quiescent act, which is acquired science, and the working act, which is the exercise of science by he who possesses it: 'he who has science thinks' (θεοϱοῦν γὰϱ γίγνεται τὸ ἔχον τὴν ἐπιστήμεν – 417 b 6). The second is a special process, different from the first which is from ignorance to science (said in its own way to be an alteration), but it presents itself as an increase in itself and in the act: 'For it is by exercise of knowledge that the possessor of knowledge becomes such in act: and this is not an alteration – for the thing develops into its own perfection and act' (εἰς αὐτὸ γὰϱ ἡ ἐπίδοσις καὶ εἰς ἐντελέχειαν – b 7). And this must be another kind of act and thus another kind of process; certainly not a process from potency to active potency, but from act to act. Here, then, the dynamic of acting expresses the intertwining of the act as first act and the second act, which is the point of departure of the metaphysical approach of St. Thomas.

[11] Aristotle, *Metaph.*, V, 7 and 12; and IX, 1-10.

Indeed, being, the mode of being, is revealed by operating, that is to say by the mode of operating. Thus from the point of view of the *via inventionis* one can say: *esse sequitur operari*. If we decentre, therefore, the activity of man towards the bottom and towards the top of each one of us, we find with a base of being, which is potent and effective, a first act for Aristotle that is not immersed in matter and is of a different kind from the rest of nature. Each man has the capacity to act according to what he is, and thus if our actions attest to the just, good and the true, it is necessary that the being of this capacity that works spiritually, which makes man in part a being heterogeneous with nature, is a being (*esse, actus essendi*) that has an emergent form above corporeal matter and not dependent on the body or the composite. Thus this being belongs inseparably to the intellective soul, like the rotundity of a circle. The human soul is a 'subsistent form' because it has the being in itself that transmits to the body and conserves it in itself when the body with death is no longer able to receive the life of the soul. The reasoning of St. Thomas is rather convincing: 'the most perfect of forms, the human soul, which is the end of all natural forms, has an activity that goes entirely beyond matter, and does not take place through a corporeal organ; namely, understanding. And because the actual being of a thing is proportioned to its activity, as has been said, since each thing acts according as it is a being (*ens*), it must be the case that the actual being of the human soul surpasses corporeal matter, and is not totally included in it, but yet in some way is touched upon by it. Inasmuch, then, as it surpasses the actual being of corporeal matter, having of itself the power to subsist and to act, the human soul is a spiritual substance; but inasmuch as it is touched upon by matter and shares its own actual being with matter, it is the form of the body'.[12]

This appears clearly even if one considers the specific activity capable of developing the human being. The perfection of understanding and willing as such lies in the possession of what is understood as intelligible in the intellect and what is loved as love in he who loves. It corresponds therefore to the human capacity to have a potentiality such as to be proportionate to

[12] '*Perfectissima autem formarum, id est anima humana, quae est finis omnium formarum naturalium, habet operationem omnino excedentem materiam, quae non fit per organum corporale, scilicet intelligere. Et quia esse rei proportionatur eius operationi, ut dictum est, cum unumquodque operetur secundum quod est ens; oportet quod esse animae humanae superexcedat materiam corporalem, et non sit totaliter comprehensum ab ipsa, sed tamen aliquo modo attingatur ab ea. In quantum igitur supergreditur esse materiae corporalis, potens per se subsistere et operari, anima humana est substantia spiritualis; in quantum vero attingitur a materia, et esse suum communicat illi, est corporis forma*' (St. Thomas Aquinas, *De spiritualibus creaturis*, 2).

the taking on of intelligible and lovable reality. Now, 'the potency of prime matter is not of this sort, for prime matter receives form by contracting it to the individual being. But an intelligible form is in the intellect without any such contraction; for thus the intellect understands each intelligible as its form is in it. Now the intellect understands the intelligible chiefly according to a common and universal nature, and so the intelligible form is in the intellect according to its universality (*secundum rationem suae communitatis*). Therefore, an intellectual substance is not made receptive of form by reason of prime matter, but rather through a character which is, in a way, the opposite (*sed magis per oppositam viam*)'.[13]

Dual Act and Dual Potency from the Operative to the Ontological

And more precisely, in analogy with the statement of Aristotle about the dual potency and dual act in the analysis of the human praxis of knowledge, both as habit and as *theoresis*, St. Thomas in going towards the depths of our

[13] *Ibidem*, 1 co. In a parallel passage of *The Treatise on Separate Substances*, the Angelic Doctor expressed himself in an analogous way: 'The matter of corporeal things, however, receives the form in a particular way, that is, not according to the common nature of form. Nor does corporeal matter act in this way insofar as it is subject to dimensions or to a corporeal form, since corporeal matter receives the corporeal form itself in an individual way. Accordingly, it becomes clear that this befits such a matter from the very nature of the matter which, since it is the lowest reality, receives form in the weakest manner; for reception takes place according to the mode of the receiver. Thereby matter, by receiving that form in a particular way, falls short in the greatest degree of that complete reception of form which is according to the totality of the form. Now it is clear that every intellectual substance receives the intellected form according to its totality, or otherwise it would not be able to know it in its totality. For it is thus that the intellect understands a thing insofar as the form of that thing exists in it. It remains therefore that if there be a matter in spiritual substances, it is not the same as the matter of corporeal things, but much nobler and finer, since it receives form according to its totality' i.e. '*Materia autem corporalium rerum suscipit formam particulariter, idest non secundum communem rationem formae. Nec hoc habet materia corporalis inquantum dimensionibus subiicitur aut formae corporali, quia etiam ipsam formam corporalem individualiter materia corporalis recipit. Unde manifestum fit quod hoc convenit tali materiae, ex ipsa natura materiae, quae quia est infima, debilissimo modo recipit formam: fit enim receptio secundum modum recipientis. Et per hoc maxime deficit a completa receptione formae, quae est secundum totalitatem ipsius particulariter ipsam recipiens. Manifestum est autem quod omnis substantia intellectualis recipit formam intellectam secundum suam totalitatem; alioquin eam in sua totalitate intelligere non valeret. Sic enim intellectus intelligit rem secundum quod forma eius in ipso existit. Relinquitur igitur quod materia, si qua sit in spiritualibus substantiis, non est eadem cum materia corporalium rerum, sed multo altior et sublimior, utpote recipiens formam secundum eius totalitatem*' (*De substantiis separatis*, c. 7)

self itself can speak about a dual act and a dual potency that are ontological: 'in composite things there are two kinds of act and two kinds of potency to consider. For first of all, matter is as potency with reference to form, and the form is its act. And secondly, if the nature is constituted of matter and form, the matter is as potency with reference to existence itself, insofar as it is able to receive this. Accordingly, when the foundation of matter is removed, if any form of a determinate nature remains which subsists of itself but not in matter, it will still be related to its own existence as potency is to act. But I do not say, as that potency which is separable from its act, but as a potency which is always accompanied by its act'.[14]

An explorer of these metaphysical sublimes, St. Thomas manages to affirm something that is surprising as regards the very high dignity of the human being: 'we see a certain gradation of infinity in things. For a material substance is finite in a two-fold manner, namely, on the part of the form which is received in matter and on the part of the "to be" itself, in which it shares according to its own mode, as being finite from below and from above. A spiritual substance – the Angel and the human soul –, however, is finite from above, inasmuch as it receives "to be" from the First Principle according to its proper mode; it is infinite from below, insofar as it is not received in a [material] subject. But the First Principle, God, is infinite in both'.[15]

[14] '*In rebus compositis est considerare duplicem actum, et duplicem potentiam. Nam primo quidem materia est ut potentia respectu formae, et forma est actus eius; et iterum natura constituta ex materia et forma, est ut potentia respectu ipsius esse, in quantum est susceptiva eius. Remoto igitur fundamento materiae, si remaneat aliqua forma determinatae naturae per se subsistens, non in materia, adhuc comparabitur ad suum esse ut potentia ad actum: non dico autem ut potentiam separabilem ab actu, sed quam semper suus actus comitetur*' (*De spiritualibus creaturis*, a. 1 co.).

[15] '*Sic igitur apparet gradus quidam infinitatis in rebus. Nam materiales substantiae finitae quidem sunt dupliciter: scilicet ex parte formae, quae in materia recipitur, et ex parte ipsius esse, quod participat secundum proprium modum, quasi superius et inferius finita existens. Substantia vero spiritualis est quidem finita superius, inquantum a primo principio participat esse secundum proprium modum; est autem infinita inferius, inquantum non participatur in subiecto. Primum vero principium, quod Deus est, est modis omnibus infinitum* (*De substantiis separatis*, c. 8). Cf. also the general principle: 'According to the Philosopher (*Phys.* ii) there is an order of precedence even in formal causes: so that nothing prevents a form resulting from the participation of another form: and thus God who is pure being, is in a fashion the species of all subsistent forms that participate of being but are not their own being' i.e. '*secundum philosophum, etiam in causis formalibus prius et posterius invenitur; unde nihil prohibet unam formam per alterius formae participationem formari; et sic ipse Deus, qui est esse tantum, est quodammodo species omnium formarum subsistentium quae esse participant et non sunt suum esse* (*De Pot.*, q. 6, a. 6 ad 5).

The Metaphysical Unification of Being and Human Activity

Now Averroes and many neo-Aristotelians, who in some sense anticipate Cartesian modernity, differently from Avicebron and the neo-Augustinian theologians, agree on the idea that thinking is the action of a spiritual substance, but they deny that this is united to the body as its substantial form. St. Thomas, instead, begins from the incontestable fact that this human praxis of thinking and willing is achieved by every man and every woman as individuals: '*hic homo singularis intelligit, vult, amat*'.[16] The '*cogito*' – in inverted commas – of St. Thomas does not finish in the separate intellect of Averroes and not even in the impersonal transcendental of the Kantian self like the modern *cogito* of Descartes because 'no one can assent to the thought that he does not exist. For, in thinking something, he perceives that he exists'.[17] There is an inseparable belonging between the thinking and the being of each human being. If, then, understanding is to the advantage of every human being and 'as an individual man', 'because it is obvious that understanding belongs to "this particular man" (as, for instance, Socrates or Plato)', one should say that it proceeds from a present principle that determines the human being as rational nature. This principle is the spiritual soul which is thus the substantial form of the human being: 'Accordingly, it must be the case that the principle of that activity which is understanding should be in "this man" in the way of a form. Now the principle of this activity is not a form whose actual being is dependent on matter and tied down to or immersed in matter, because this activity is not effected by means of the body, as its proven in III *De Anima* [4, 429 a 24]; and hence the principle of this activity possesses an activity that has nothing in common with corporeal matter. Now, the way in which each thing acts is a consequence of its being. Hence the actual being of that principle must be an actual being which is raised above corporeal matter and not dependent on it. Now this is characteristic of a spiritual substance. It is necessary to say, therefore, if the preceding considerations are put together, that some kind of substance is the form of the human'.[18]

[16] The principle *hic homo singularis intelligit* is repeated 14 times, cf. *De unitate intelectus contra Averroistas*, capp. 3 e 4. For *vult* cf. *De Malo*, q. 6 art. un.

[17] '*Nullus potest cogitare se non esse cum assensu: in hoc enim quod cogitat aliquid, percipit se esse*' (*De veritate*, q. 10, a. 12 ad 7).

[18] '*Oportet igitur principium huius operationis quod est intelligere, formaliter inesse huic homini. Principium autem huius operationis non est forma aliqua cuius esse sit dependens a corpore, et materiae obligatum sive immersum; quia haec operatio non fit per corpus, ut probatur in III de anima; unde principium huius operationis habet operationem sine communicatione materiae corporalis. Sic*

Having rejected this *Averroist* opinion as impossible, St. Thomas takes into consideration that opinion that most seduced Patristic thinkers and the neo-Augustinians theologians who, following Plato, argued that the concrete individual thinks but nonetheless the spiritual substance is not united to the body as its form but has the same function as that of a sailor for a ship. Now, if the soul were not the form of the body, it and its parts would not obtain from the soul its specificity and identity, which appears evidently false, because, on separating from the soul the eye, the brain, the heart, the flesh and bone can longer be said to be such 'except equivocally, like an eye in stone or in a picture'.[19]

This metaphysical union of human activity, and thus of individuality and personality themselves constitutes a 'new event in thought', that is to say an absolute innovation in Christian thought that was unknown to Patristic thought, which St. Thomas managed, however, to develop thanks to the principle of the ontological continuity of the species of the Pseudo-Dionysius '*Supremun infini attingit infimum supremi*'. Thus: 'the human soul, which is the lowest in the order of spiritual substances, can communicate its own actual being to the human body, which is the highest in dignity, so that from the soul and the body, as from form and matter, a single being results'.[20] And, concluding in a masterful way, St. Thomas demonstrates to his neo-Augustinian colleagues of the Faculty of Theology, with an extraordinary beat of his intellectual wings, the defeat to which their theory exposed them in relation to the followers of Averroes of the Faculty of Philosophy: 'But if a spiritual substance were composed of matter and form, it would be impossible for it to be the body's form: because it is essential to matter that it be not in anything else, but that it should itself be the primary subject'.[21]

autem unumquodque operatur secundum quod est; unde oportet quod esse illius principii sit esse elevatum supra materiam corporalem, et non dependens ab ipsa. Hoc autem proprium est spiritualis substantiae. Oportet ergo dicere, si praedicta coniungantur, quod quaedam spiritualis substantia, sit forma humani corporis' (*De spiritualibus creaturis*, a. 2 co.).

[19] Aristotle, *De Anima*, II, 1, 412 b 20 f.

[20] '*Attingitur autem a materia corporali ea ratione quod semper supremum infimi ordinis attingit infimum supremi, ut patet per Dionysium VII cap. de Divin. Nomin.; et ideo anima humana quae est infima in ordine substantiarum spiritualium, esse suum communicare potest corpori humano, quod est dignissimum, ut fiat ex anima et corpore unum sicut ex forma et materia* (*De spiritualibus creaturis*, a. 2 co).

[21] '*Si vero substantia spiritualis esset composita ex materia et forma, impossibile esset quod esset forma corporalis: quia de ratione materiae est quod non sit in alio, sed quod ipsa sit primum subiectum*' (*loc. cit.*). Averroism, together with *Alexandrinism*, were expressly condemned by the Fifth Lateran Council under Leo X with the Bull *Apostolici regiminis* (1513): 'Now,

Freedom of Will and the Divine Conatus

It appears that during this same period St. Thomas was the first to grasp the corollary that was most destructive of the Averroist position on the basis of which the human being does not specifically have free choice in his acts but, rather, his will is moved to choosing out of necessity, although it is not subjected to coercion. Indeed, not every necessary thing is violent but only that which has an external principle, and it follows from this that the will is necessarily moved without violence by an internal principle such as the intellect. For St. Thomas 'this opinion is heretical. For it takes away the reason for merit and demerit in human acts, as it does not seem meritorious or demeritorious for persons to do necessarily what they could not avoid doing. It is also to be counted among the oddest philosophical opinions, since it is not only contrary to faith but also subverts all the principles of moral philosophy. For if nothing is within our power, and we are necessarily moved to will things, deliberation, exhortation, precept, punishment, and praise and blame, of which moral philosophy consists, are destroyed'.[22] For St. Thomas, instead, if it is true that the intellect as a faculty of the true precedes the will and guides it, and he indeed states that '*primum principium motionis est ex intellectu: hoc enim modo bonum intellectum movet etiam ipsam voluntatem*', however in the order of the exercise of the act, which is that of the real actuating itself of the freedom of the person, the relationship is overturned and it is the will which has as an object the end, or the good, that decides the action and confers a moral quality on the exercise of the intelligence and all the other faculties and habits. St. Thomas managed to demonstrate the freedom of the will which is another new 'event of thought' or an absolute innovation in the history of philosophy and he found for it the formula: '*Intelligo enim quia volo; et similiter utor omnibus po-*

the sower of tares has dared to sow and multiply extremely dangerous errors...above all on the nature of the rational soul, according to which it is mortal or unique for all men…we condemn and rebuke all those who state that the intellective soul is mortal or unique for all men…in fact the soul is not only truly, of itself and essentially, the form of the human body…but it is also immortal, and, given the multitude of bodies in which it is individually infused, it can be, must be and is multiplied' (*Doctrine on the Soul, against the Neo-Aristotelians*, Denzinger-Huenermann, 1440, p. 621).

[22] '*Haec autem opinio est haeretica: tollit enim rationem meriti et demeriti in humanis actibus. Non enim videtur esse meritorium vel demeritorium quod aliquis sic ex necessitate agit quod vitare non possit. Est etiam annumeranda inter extraneas philosophiae opiniones: quia non solum contrariatur fidei, sed subvertit omnia principia philosophiae moralis. Si enim non sit liberum aliquid in nobis, sed ex necessitate movemur ad volendum, tollitur deliberatio, exhortatio, praeceptum et punitio, et laus et vituperium, circa quae moralis philosophia consistit*' (*De malo*, q. 6 co).

tentiis et habitibus quia volo', clearly bringing out – against any rationalist determinism – the real dominion of freedom and this of the end, which is the good, in the behaviour of a person. And here St. Thomas with sophistication that is specific to him brings out the contradiction of his Averroist colleagues by referring to the authority of their mentor, Averroes: 'And so also the Commentator in his *Commentary on the De anima* defines habit as what a person uses at will'.[23]

Now, because it is not possible to dwell on this subject *ad infinitum*, one should necessarily state that, as regards the first movement of the will, that is to say the move from potency to act or the placing in act of freedom, the will of every human being is moved by an agent by impulse of which it begins to will freely. This agent cannot be a celestial body nor anything material or of the organism such as the genes, as some affirm today, because the will is not corporeal potency. 'Therefore, we conclude', St. Thomas observes, 'as Aristotle concludes in the chapter on good fortune in the *Eudemian Ethics*, that what first moves the intellect and the will is something superior to them, namely, God'.[24] We owe to Aristotle the introduction of the divine conatus as the founding basis of human freedom, which St. Thomas read in *Eudemian Ethics* where an instinct is affirmed, that is to say a 'starting-point of motion in the soul'[25] which passes by way of Spinoza and reaches P. Ricœur.[26]

The Existential and Metaphysical Aristotelian-Thomistic Approach: the Circularity of Science and Knowing Yourself

We can recapitulate by observing that the approach of Aristotle, in this determining of the human being in the two moments of potency and act, of the first act and the acts of the faculties and habits, and of act as the habit of science and as exercise of it, followed by St. Thomas with the act of form and the act of being, is very acute, existential and metaphysical at one and the

[23] '*Unde et Commentator definit habitum in III de anima, quod habitus est quo quis utitur cum voluerit*' (*loc. cit*).

[24] '*Relinquitur ergo, sicut concludit Aristoteles in cap. de bona fortuna, quod id quod primo movet voluntatem et intellectum, sit aliquid supra voluntatem et intellectum, scilicet Deus* (*loc. cit.*).

[25] *τὸ δὲ ζητούμενον τοῦτ᾽ ἐστί, τίς ἡ τῆς κινήσεως ἀρχὴ ἐν τῇ ψυχῇ. δῆλον δὴ ὥσπερ ἐν τῷ ὅλῳ θεός, καὶ κἀν ἐκείνῳ. κινεῖ γάρ πως πάντα τὸ ἐν ἡμῖν θεῖον: λόγου δ᾽ ἀρχὴ οὐ λόγος, ἀλλά τι κρεῖττον: τί οὖν ἂν κρεῖττον καὶ ἐπιστήμης εἴη καὶ νοῦ πλὴν θεός;* (*Eth. Eudem.*, VIII, 1248 a 25 ff.). Cf. C. Fabro, 'Le liber de bona fortuna chez Saint Thomas', *Revue Thomiste*, 1988, p. 356 ff.

[26] P. Ricœur, *Sé come un altro* (Milan, 1993), pp. 429-431.

same time: existential because it comes drawn from the analysis of human praxis as increase of the life of the spirit (intelligence and freedom in St. Thomas), and at the same time metaphysical because it draws on the being as being as potency and act, and then as act in act in its foundation which is Logos and Principio at one and the same time. The essential in this anthropological legibility of being is the analogical decentring towards the bottom, that is to say the self of each human being, and the re-centring towards the top, that is to say God; this is what the late St. Thomas does: '*Deus est et tu: sed tuum esse est participatum, suum vero essentiale*' (*In Psal*. 34, 7).

Therefore, neuronal and philosophical centrality in acting and decentring in the direction of a foundation of act and potency are equally and jointly constitutive of an ontology of the human being in terms of act and potency. Therefore only the human being has this double legibility: the external objective reading, common to all the beings of nature, which is the subject of the sciences (*epistémê*), and the approach of auto-reflection, which belongs to philosophy (*sophia*), according to the Socratic precept 'know yourself', which understands being as an act of an active potency which we call the 'soul'.[27] Thus only a human being is able to create a circularity between this double legibility, seeing, so to speak, externally, the functioning of his brain with new sensors that portray it in film-like fashion, and interpreting from the inside this film-like portrayal starting from auto-reflection on himself.

There is nothing that is more ours than our brain yet there is nothing that we know less about. The ancients thought that the heart was the centre of life because it beats constantly like a pump and tells us 'I am here'.[28] On

[27] St. Thomas Aquinas, *Q. d. De Spiritualibus Creaturis*, a. 1.

[28] Indeed, St. Thomas says: '*Secundum igitur quod anima est forma corporis, non potest esse aliquid medium inter animam et corpus. Secundum vero quod est motor, sic nihil prohibet ponere ibi multa media; manifeste enim anima per cor movet alia membra, et etiam per spiritum movet corpus*' (*Q. d. De Spiritualibus Creaturis*, a. 3 co.). Also: '*unumquodque operatur in remotiora per id quod est maxime proximum. Sed vires animae diffunduntur in totum corpus per cor. Ergo cor est vicinius quam ceterae partes corporis; et ita mediante corde unietur corpori*' (*Q. d. De Anima*, a. 9, arg. 13). Also: '*cor est primum instrumentum per quod anima movet ceteras partes corporis; et ideo eo mediante anima unitur reliquis partibus corporis ut motor, licet ut forma uniatur unicuique parti corporis per se et immediate*' (*Q. d. De Anima*, a. 9, ad 13). Again, from a general point of view: '*cum anima rationalis sit perfectissima formarum naturalium, in homine invenitur maxima distinctio partium propter diversas operationes; et anima singulis earum dat esse substantiale, secundum illum modum qui competit operationi ipsorum. Cuius signum est, quod remota anima, non remanet neque caro neque oculus nisi aequivoce. Sed cum oporteat ordinem instrumentorum esse secundum ordinem operationum, diversarum autem operationum quae sunt ab anima, una naturaliter praecedit alteram, necessarium est quod una pars corporis moveatur per aliam ad suam*

the contrary, the brain was, so to speak, the great silence or the sealed box of our body.[29] Today however the brain opens itself up and shows itself, in part because of the neurosciences, as being the centre of the body, and this may turn out to be a turning point for a new beginning where external experience can be joined to internal experience and science can be joined to philosophy, each in their respective functions and consistencies and in their mutual circularity. This was not present in ancient philosophies, or in Medieval, modern or contemporary thought, and if the human being is analysed, he is analysed from a formal point of view without these dynamic and circular links with scientific knowledge and auto-reflective knowledge of my body and my brain. In truth, it is not that I am my body, not even its masterpiece, the brain: I am neither my brain nor my body; I have a brain

operationem. Sic ergo inter animam secundum quod est motor et principium operationum et totum corpus, cadit aliquid medium; quia mediante aliqua prima parte primo mota movet alias partes ad suas operationes, sicut mediante corde movet alia membra ad vitales operationes: sed secundum quod dat esse corpori, immediate dat esse substantiale et specificum omnibus partibus corporis. Et hoc est quod a multis dicitur quod anima unitur corpori ut forma sine medio, ut motor autem per medium. Et haec opinio procedit secundum sententiam Aristotelis qui ponit animam esse formam substantialem corporis. Sed quidam ponentes secundum opinionem Platonis animam uniri corpori sicut unam substantiam, alii, necesse habuerunt ponere media quibus anima uniretur corpori; quia diversae substantiae et distantes non colligantur, nisi sit aliquid quod uniat eas. Et sic posuerunt quidam spiritum et humorem esse medium inter animam et corpus, et quidam lucem, et quidam potentias animae, vel aliquid aliud huiusmodi. Sed nullum istorum est necessarium, si anima est forma corporis; quia unumquodque secundum quod est ens, est unum. Unde cum forma secundum seipsam det esse materiae, secundum seipsam unitur materiae primae, et non per aliud aliquod ligamentum' (*Q. d. De Anima*, a. 9 co.).

[29] However, St. Thomas had already acutely observed the absolute necessity, for the working of the mind, of the state of perfection of the body: '*naturale est animae quod indigeat phantasmatibus ad intelligendum; ex quo tamen sequitur quod diminuatur in intelligendo a substantiis superioribus. Quod autem dicitur, quod anima a corpore praegravatur, hoc non est ex eius natura, sed ex eius corruptione, secundum illud Sapient. IX: corpus quod corrumpitur aggravat animam. Quod vero dicitur quod abstrahit se a nexibus corporalibus ut se intelligat, intelligendum est quod abstrahit se ab eis quasi ab obiectis, quia anima intelligitur per remotionem omnis corporeitatis; non tamen ab eis abstrahitur secundum esse. Quinimmo, quibusdam corporeis organis laesis, non potest anima directe nec se nec aliud intelligere, ut quando laeditur cerebrum*' (*Q. d. De Spiritualibus Creaturis*, a. 2 ad 7). Also: '*Hanc igitur oportet esse dispositionem corporis cui anima rationalis unitur, ut scilicet sit temperatissimae complexionis. Si quis autem considerare velit etiam particulares humani corporis dispositiones, ad hoc inveniet ordinatas, ut homo sit optimi sensus. Unde, quia ad bonam habitudinem potentiarum sensitivarum interiorum, puta imaginationis et memoriae, et cogitativae virtutis, necessaria est bona dispositio cerebri. Ideo factus est homo habens maius cerebrum inter omnia animalia, secundum proportionem suae quantitatis; et ut liberior sit eius operatio habet caput sursum positum; quia solus homo est animal rectum, alia vero animalia curva incedunt*' (*Q. d. De Anima*, a. 8 co.).

and a body but – as I have tried to show – in order to understand my 'being' I must know what to have a brain means, to have a body means, through that knowledge of them that experience and science offer to me.

Philosophy follows its own synthetic method: it acts with the experimental data provided by science and neuroscience and the principles of reason but moves them within the transcendent reality of the soul as a spiritual free subject[30] and of God the Creator. Thus experience, science and philosophy are fused in their respective functions and consistencies and a 'breach' of movement is made towards the limit that always keeps the consciousness of a person alert and vigil.

[30] The fact that sensitive knowledge precedes intellectual knowledge in the human being, the sensitive origin of human intellectual knowledge and the affirmation that the soul (the profound self of each of us) can come to know itself as spiritual only through the intellectual species that are abstract from the sensitive one, have prevented most of the time not only the understanding but also the actual reading of the texts of St. Thomas who focuses on the real issue in question and shows that "the principle of human knowledge comes from sense. However, it is not necessary for everything that man knows to be submitted to sense or that it is immediately known only by means of a sensitive effect". Indeed, he affirms what we may call the decisive epistemological position of the Socratic principle of "know yourself": "The very intellect knows itself by means of its own act, which is not submitted to sense. In the same way, it also knows the interior act of will, since will is somewhat moved by the intellectual act and since intellectual act is caused in another way by will, like the effect is known by means of the cause and the cause by means of the effect" i.e. "principium humanae cognitionis est a sensu; non tamen oportet quod quidquid ab homine cognoscitur, sit sensui subiectum, vel per effectum sensibilem immediate cognoscatur; nam et ipse intellectus intelligit seipsum per actum suum, qui non est sensui subiectus: similiter etiam et interiorem actum voluntatis intelligit, in quantum per actum intellectus quodammodo movetur voluntas, et alio modo actus intellectus causatur a voluntate, ut dictum est, sicut effectus cognoscitur per causam, et causa per effectum" (*De Malo*, q. 6, a. un. ad 18). This is a decisive point because St. Thomas also states that 'we would not be able to obtain knowledge about separate intellectual substances either though reason or through faith, unless our soul knew on its own to be an intellectual being': '*Cum enim de substantiis separatis hoc quod sint intellectuales quaedam substantiae cognoscamus, vel per demonstrationem vel per fidem, neutro modo hanc cognitionem accipere possemus nisi hoc ipsum quod est esse intellectuale, anima nostra ex seipsa cognosceret*'" (*Summa contra Gentiles*, III, 46). Thomas also accepts that is it because of the spiritual soul that the human intellect can raise itself to God: 'the soul itself, through which the human intellect ascends to knowledge of God': *'etiam ipsa anima per quam intellectus humanus in Dei cognitionem ascendit*' (*Ib*., I, 3).

Sources of Human Comprehension and Incomprehension

Are There Innate Mechanisms That Make Us Social Beings?

Uta Frith[1]

Introduction

We humans consider pride ourselves in being an ultra-social species with a strong desire to learn from each other and to cooperate with each other (Boyd & Richerson, 2009; Tomasello & Vaish, 2012). Folk psychology lets us believe that we have conscious control over our most treasured social abilities, such as empathy, fairness and morality, and that we pass them on to the next generation through teaching normative rules.

However, this folk belief does not fit with the compelling demonstration that we are not conscious of most of our cognitive and social abilities, and that these are already in place in early childhood well before normative rules are taught (Tomasello, 2008). Furthermore, social abilities of an apparently high level of complexity, such as altruistic helping, mind-reading, and reputation management, can be observed in non-human animals including fish, insects and birds (Frith & Frith, 2012). In the present paper, I will reflect on where our social abilities come from, how they are organised and what happens when they go wrong.

At present we struggle to understand the putative presence of conscious and unconscious systems in the mind. Daniel Kahneman (2011) elaborated on these systems, economically dubbing them I and II, or fast and slow. According to Kahneman, the unconscious system I is fast and powerful and rules our mental lives far more than we realise. However, it does submit occasionally to reasoning as presented by system II. The conscious system II is slow, weak and also more error prone than we realise. It provides justifications and rationalisations for behaviour that is actually caused by unconscious processes. I will argue that the fast system is based on *instincts*, and that this is where we should look for innate social mechanisms.

The problem with instincts

The term instinct has had an image problem, and understandably so. There are two main obstacles to using this unfashionable term. First, instincts

[1] UCL Institute of Cognitive Neuroscience and Interacting Minds Centre, University of Aarhus.

conceived as a sequence of rigidly hard-wired behaviour are incompatible with contemporary ideas of how the brain works. The very word hard-wired contradicts the idea of brain plasticity. Second, instincts are often defined as excluding learning, when learning is in many ways the essence of what the brain does. We learn all the time and adapt our brain to new conditions.

I propose to use the term 'start-up kit' to suggest that there is both a given predisposition and learning. Learning is needed to realise and to tune up any mental ability, however innate. I will continue to use the term innate, even though, like instinct, this also has had a bad press. Some difficult obstacles have to be overcome. For instance, it is inimical to emphasise inequality that is inherent in the notion of genetically based abilities and disabilities and smacks of genetic determinism. However, genetically based individual differences are undeniable, while genetic effects are probabilistic and act in interaction with environmental and social factors. Strong environmental effects can sometimes trump genetic effects. For example, genetically caused disorders are amenable to intervention, and gene therapy has been shown to be successful (Sheridan, 2011).

It is generally agreed that the brain is a prediction machine (Knill & Pouget, 2004). The prediction of the social behaviour of other agents is of as much value and importance as the prediction of events in the physical world. There is much less controversy in the assumption that the brain comes equipped with innate start-up kits for predicting the physical world. To mention just two examples: the ability to detect of numerical magnitude depends on an intact intra-parietal sulcus (Piazza *et al.*, 2007), and the ability to navigate in space depends on an intact hippocampal formation (Burgess & O'Keefe, 2011). It is encouraging that specific neural mechanisms have been identified for several cognitive processes, and it is highly likely that this is true also for the social mind/brain. But how can we know there are circumscribed social mechanisms, and how would we test the hypothesis that they rest on innate start-up kits? It is research on the cognitive basis of autism, which gives me strong reasons for this belief.

What do 'start-up kits' do?

The argument can be made that powerful learning mechanisms are sufficient to explain social behaviour. Social learning is likely to be the main driver for human adaptation over generations (Boyd, Richerson & Henrich, 2011; Heyes, 2012). If learning is so powerful, why require innate mechanisms at all? I believe we need them to explain social learning within a single brain. For one thing, life is too short to learn everything that can be learned, if there are no predispositions to set priorities. For another thing,

not everything can be learned equally easily and this is hard to explain if learning mechanisms don't have prior dispositions. For instance, fear of snakes can be learned quickly, but not fear of flowers (Cook & Mineka, 1989). Learning proceeds within remarkably strict limits. These are often referred to as constraints. However, I believe it is time to think of innate predispositions not as constraining learning, but as enabling learning.

It is also important to get rid of the idea that innate mechanisms must be present at birth and look for them only in young infants. Indeed there are some time bombs, which are set to detonate at different stages of life, e.g. sexual maturity and child bearing. Evolution has resulted in adaptive mechanisms, manifest in the way the brain is organised, and in the way learning gets off to a quick start when it is needed. This can happen at different ages.

For the purposes of this paper I assume that start-up kits are genetically programmed predispositions for specific computational processes, located in different circuits of the social mind/brain. They enable fast track learning in vital domains. This learning can be seen as a necessary calibration of the mechanism for given environmental conditions, with recalibration when circumstances change sufficiently. The idea is that there is a smoothly working mechanism that responds automatically to the right stimuli, like lock and key.

Autism as a crucible

Autism Spectrum Disorder (ASD) is a heterogeneous collection of disorders caused by a multitude of genetic and epigenetic causes (Geschwind, 2009). In view of this heterogeneity the fact that a clinical diagnostic category of autism exists is remarkable. One way to explain this coexistence of multiplicity at one level and unity at another level, is that all distal causes converge in a bottleneck. This bottleneck is the social brain, which develops atypically (Abrahams & Geschwind, 2010; Pelphrey *et al.*, 2011). At another level, the causal paths diverge again to give rise to impairments in a variety of social behaviours. These behaviours are also affected by other distal factors, such as education and learning, which alleviate or aggravate the condition. It is easy to picture this in a three-level framework (Morton & Frith, 1998) as shown in Figure 1. I believe that it is by investigating the nature of the mechanisms at the cognitive level that the anatomy of the social mind/brain can be laid bare.

I will focus here on a particular cognitive function, Theory of Mind, also referred to as ToM or mentalising (for a recent review see Apperley, 2012). I would place '*mentalising failure*' in the big oval space in the middle of Figure 1, with smaller ovals representing other hypothetical cognitive dysfunctions. The mentalising hypothesis claims to apply to all individuals on the autistic

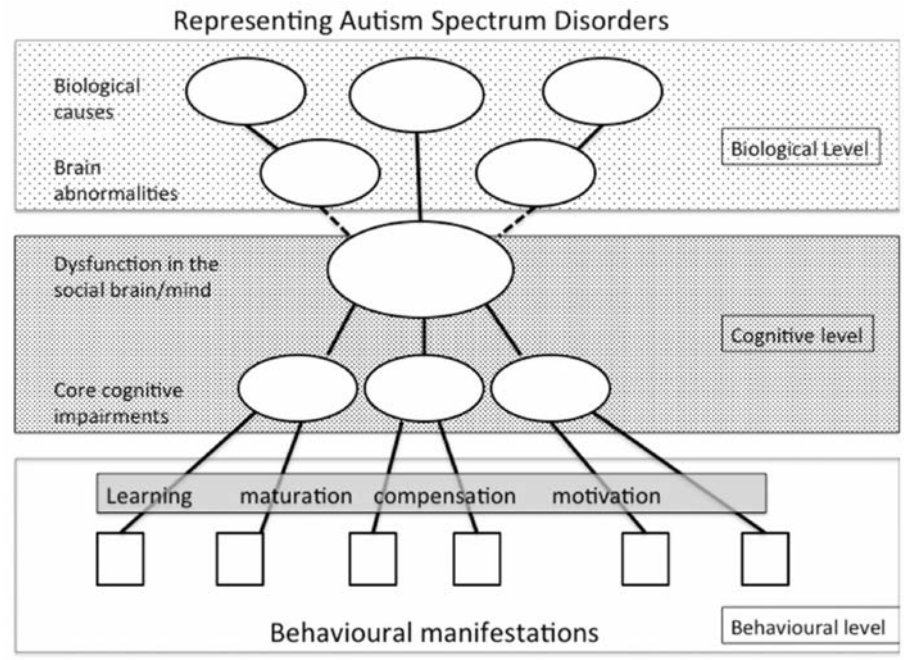

Figure 1. Three-level framework.

spectrum, regardless of what additional cognitive problems they might have. Empirical studies over the last 20 years have given some credence to this claim. They have shown that autistic individuals show atypical brain activity in the mentalising system of the social brain. They have also shown that their characteristic social impairments can be parsimoniously explained via mentalising failure (for a recent review see Frith, 2012). However, there may be additional dysfunctions that act as proximate causes of the social impairments seen in autism (for a recent review see Chevallier *et al.*, 2012).

Gotts *et al.* (2012) derived a summary picture of the social brain as revealed by autism. This is shown in Figure 2 (see p. 318). The brain regions indicated in red are active when volunteers perform various social tasks in the scanner. For instance, they are active when a task involves mentalising, but not active, when it does not involve mentalising. The difference between these two conditions is what appears as red blobs in typical brain images. Over and over again the blobs have been found in these regions: medial prefrontal cortex, superior temporal sulcus, temporoparietal junction, temporal poles, basal temporal regions, inferior frontal gyrus, amygdala, pre-

cuneus and anterior insula. Together the regions form a network that we can call the social brain. Of course, to understand mental mechanisms it is not sufficient to know their location in the brain. However, the evolutionary history of the brain regions pinpointed can give clues to the origins of the mechanisms and how they became incorporated into genetic programmes.

A metaphor for the social mind/brain

I am going to use the visual prop of a house as a structure representing the human social mind/brain. I will act as an estate agent to tempt you to invest in this substantial property. I know you are not going to move in, it is after all a metaphor, but you may enjoy viewing it. I will guide you through its five floors and tell you about its ancient history. Two firms inhabit the property, Kahneman I and II. The bulk of the house belongs to the firm of Kahneman I; only the upper floor and the attic belong to the firm of Kahneman II.

The most important thing to know about the house is that it is a radio station. The daily business of the inhabitants is broadcasting and receiving news. When you view the house you can experience how utterly enthralled the inhabitants are by everything that is going on in other houses. You can also see that decisions have to be made constantly to switch on or off various listening and broadcasting devices. As a scrupulous estate agent, I will not conceal that some of the switches and information processing devices are liable to dysfunction. In extreme cases, parts of the firms may close down. This is not entirely surprising given the highly complex nature of the information they trade.

Here is a quick look at the floor plan followed by more detailed viewing.

Basement – Recognising other agents
Ground floor – Being part of a group
Bel-étage – Taking account of others' mental states
Top Floor – Intelligence services: explicit Theory of Mind
Attic Tower – The Self

Basement – Recognising other agents

The house has sound foundations. The cavernous basement contains ancient structures. This is tried and trusted machinery that never fails. It is shared with virtually all other animals, particularly mammals and is and deeply unconscious. This machinery allows us to recognise social agents, kin and non-kin, and moreover, individual others. Survival depends on ultra-fast learning on which agent to approach and which to avoid. These

mechanisms are in excellent order and suggest an innate predisposition to prioritise other agents over any other stimuli in the environment. This may be because social stimuli are intrinsically rewarding (Delgado, 2007; Krach *et al.*, 2010). The reward yields dividends in affiliation and love, but also in gaining status and winning competitions.

To help the visitor, there is a video in a roomy vault, which explains how mammals bond with their young to suckle and nurture them. This may have provided a powerful push towards a social capacity in the human brain that we refer to as love. We can glance into a lab producing potions made from hormones such as oxytocin, vasopressin and testosterone. The effects of the potions go from nurturing the young to sexual interaction and pair bonding to fierce parochialism and competition (Carter *et al.*, 2008; McCall &Singer, 2012).

Another video shows fighting deer with huge antlers. Here you can see a mechanism in action that can evaluate individual differences in prowess. This may also be relevant to recognition of dominance hierarchies. We go back to the first video and see that already at the stage of suckling, there is differentiation between siblings, so that some get a better place at the milk producing glands than others. It is not only rewarding to be near the top of the hierarchy, but it can be fatal to be at the bottom (e.g. Petticrew & Davey-Smith, 2012).

A slide show gives an overview over the ancient evolutionary roots of the machinery and we can see that this machinery may primarily serve selfish aims. We see flashing up the phrase 'nature red in tooth and claw' and realise that this applies to social lives in particular. The mechanisms come as integral parts of the house, but they improve by learning. Learning is not left to chance, but is guided by strong rewards. Rewards are strongly linked to affiliation, sexual reproduction, nurturing, having status and winning over rivals (e.g. Zink & Meyer-Lindenberg, 2012). It would seem difficult to subdue these robust mechanisms and the strong emotions they give rise to, and yet it is possible to do so. They can be silenced temporarily using conscious inhibition. This type of control, using cultural rules and moral guidelines is the job of the firm Kahneman II, at the top floor of the house.

Ground floor – Being part of a group

It is party time. If you are a gregarious extravert, you will love this floor. The mechanisms here are of particular value to animals whose survival depends on living in groups. These mechanisms here too are robust and well honed over millennia of evolution. They rarely fail, and if you are an introvert, you can always find an off switch. Still, you should hesitate to use this switch. The mechanisms can make life very rewarding for you. They foster

alignment to other members of your group. The result is a We-mode of social perception (Tuomela, 2007). Turn on the We-mode and you share experiences with others. In the basement, the I-mode is more often in use. The basement takes account of others in terms of how well they serve the Self, but this egocentric mode is bad manners here on the ground floor. You are enveloped and submerged in a greater whole. There is dance music with inviting rhythms; there is choral singing and in one room the Ode to Joy can be heard in the background: "Seid umschlungen Millionen…"

A video of a football game is on show, with the camera turned mainly on the crowd who chant and move as one. Another video shows herds of animals in the African Savannah. It all makes sense when you think about the ground-floor mechanisms as herd instinct (Raafat, Chater & Frith, 2009). Another video shows you that the alignment in motion and emotion also supports empathy (Singer & Lamm, 2009). But there are more complex forms of empathy, (Blair, 2005), and Kahneman II is in charge of them on the top floor.

You cannot be alone on this floor for long. On every video being demonstrated you see other people. The videos impress on you that human beings, like other social animals, have a strong tendency for compliance and conformity (Cialdini & Goldstein, 2004). They want to be where other human beings are. Doing as others do is rewarding and makes you feel benevolent towards those less fortunate than you (van Baaren *et al.*, 2004). In one room you are invited to play a game on a slot machine set up in a virtual web with other players. You soon find that a good way of winning is by copying others (Rendell *et al.*, 2010). Social learning pays. One neat mechanism works in such a way that eye gaze follows eye gaze. This makes social learning very simple: you automatically follow the gaze of another person.

There are some dark rooms, but I will show them to you. Here we can glimpse oppositional behaviour, and also aggression and violence. We see that an individual who behaves out of line with the group and sticks out, risks ostracism. Even the bare hint of ostracism makes the mechanism that promotes copying behaviour work harder (Over & Carpenter, 2008). While it is rewarding to conform, it is also painful not to conform (Haun, van Leeuwen & Edelson, 2012). A video shows that strong tendencies to do what others do result in social network formation (Christakis & Fowler 2012). They can transmit good feelings as well as bad; and they are responsible for all sorts of ideas spreading quickly. The analogy with infectious diseases is not far-fetched. Still, control systems can stop the spread.

More rooms with more video shows can be visited, as this is a very large floor. We can take part in a quick experiment. You are asked to imagine

yourself in an emergency. You have a split second to decide which other people you would turn to. The odds are that these others are similar to you. This automatic tendency is also the basis of social prejudice. We can do another experiment to show how instantly we can classify others into an in-group or outgroup. It takes only the slightest prompts (for a review see Shkurko, 2012). A horror movie about Zombies shows how frighteningly easy it is for human beings of all ages to turn off prosocial group oriented behaviour, and become aggressive and nasty to members of an out-group (Bernhard, Fischbacher & Fehr, 2006).

There is so much to see on this floor, so many emotions to feel, that while hurrying through we can only get a glance at what regulates group behaviour. In human societies there is an abundance of culturally evolved rules and regulations, some enshrined in law. Here Kahneman II is in charge. But the ground floor has its own built-in rules. These are the same that play out in automatic fashion throughout the animal kingdom. For instance, a tit-for-tat strategy is common across many social species. We recognise the almost uncontrollable feeling of revenge, but this mechanism counteracts the invasion by free riders (Raihani & Bshary, 2011). Troubleshooting like this is something that the basically altruistic ground floor often has to do. How come it is so altruistic? One fundamental drive in many social species is an egalitarian drive, whose evolutionary roots have been modelled (Gavrilets, 2012). Hence, fairness is rewarding (Tabibnia & Lieberman, 2007) and inequity is aversive. Aversion to inequity can induce individuals to punish selfish culprits even if it is at a cost to themselves (Raihani & McAuliffe, 2012). A brief slide presentation of neighbouring properties shows us that not all houses are the same. There are a minority of individuals who value their own selfish interest much more than equity (Haruno & Frith, 2010; Morishima *et al.*, 2012). Group oriented individuals too can become free riders, particularly if they see other free riders doing it and getting away with it too.

It is difficult to see inside the hub of the ground floor machinery: all sensory modalities have a role to play, the movement, sound, smell and touch of others, but visual cues are particularly important. Many believe that the real engine here is the mirror neuron system. Located in the inferior frontal gyrus and the intraparietal regions, this serves as a neural basis for the links between perception and action (Kilner, Friston & Frith, 2007). Innate mechanisms might set up the capacity to have a mirror neuron system but processes such as association learning are also likely to be involved (Heyes, 2011). Perhaps the most amazing thing about the hub is that the same mechanism applies to glue together perception and action within one and the same person as across persons. This truly is a We-mode engine.

You might ask whether the ground floor mechanisms are exclusively in the service of the We-mode or whether they can they also work in the I-mode. I think they can. For instance, fairness tends to be monitored within in-groups, and thus implicitly compares individuals (Fehr, Bernhard & Rockenbach, 2008). Altruism can be seen as the easy thing to do in the We-mode, or the effortful thing to do in the I-mode. A general house rule is: save energy and always do the easy thing.

The understanding of group hierarchies is another crucial aspect of group living: you need to know your place (Zink *et al.*, 2008). We have already met examples of mechanisms that underlie dominance and status in the basement. This was when they involved an egocentric perspective. However, status can also be a group concern. It seems that rapid switching between I-mode and We-mode is part of our social nature and there are clear benefits to be had from switching. There is direct reciprocity and the I-mode is sufficient here: you scratch my back and I scratch yours. But there is also indirect reciprocity: you scratch my back, and somebody else will scratch your back, or perhaps your friend's back (Nowak & Sigmund, 1998). This demands a We-mode.

I have to point out that things sometimes do go wrong on the ground floor. In the case of autism, suggestions have been made for a number of the mechanisms located here to be faulty. Processes underlying eye gaze following have a reasonably strong claim of being critically abnormal in autism, and there are problems in spontaneously copying others and in experiencing emotional contagion. Autistic people stick out, and this leads to ostracism, for instance in the form of bullying or neglect. Anecdotal observations suggest that many autistic people actually like to be where other people are; many are keen, even desperate, to have friends. They join clubs and maintain web-based support groups. They learn from observing other people; they imitate them, if sometimes clumsily. They pick up on social stereotypes and can make in-group – out-group distinctions (Hirschfeld, Bartmess, White & Frith, 2007). For instance they like to distinguish themselves from NTs, or neurotypical people and are proud to be different. There is then a strong possibility that the mechanisms on this floor of the house are dissociable and that autism provides a way to dissociate them.

Bel-étage – Taking account of other's mental states

This floor is a big selling point of the house. You can walk through some spacious public rooms, set up ready for visitors such as you. Here is the place of a rather more delicate mechanism, not as ancient as the ones on the lower floor and more exclusive and also more fragile. In fact this machinery is rare

in the animal kingdom. It makes humans social in a rather special way. It senses and processes information at ranges that are invisible to sensory modalities: It automatically takes into account your point of view, your desires, your intentions and beliefs, in short, your mental states that you might have thought were entirely private. In fact they give invaluable clues as what you are going to do next and that is what the information processing machines on this floor are interested in. They can read your wishes before you even realise you have them and can extrapolate from your beliefs what events will surprise you. This ability is known as *implicit mentalising*. It is a We-mode thing.

This automatic mind reading mechanism is well worth having. It saves time, effort and errors, because it spares you from having to make complex conscious inferences just to know that someone who has not seen an event will not know about it. You see a demo where you watch a little boy who believes that an empty box contains sweets: you can instantly anticipate that he will be disappointed when looking in the empty box.

A slide show explains that Apperly & Butterfill (2009) argue for two systems to track beliefs and other mental states, one implicit, and one explicit. Implicit mentalising is present in human babies (Onishi & Baillargeon, 2005; Kovacs, Teglas & Endress, 2010) and in some non-human animals (Emery & Clayton, 2009; Bugnyar, 2011). The explicit model is on the top floor, belonging to Kahneman II. The idea of two models chimes in with the finding that in autism implicit mentalising remains faulty even in very able adults who have acquired excellent skill in explicit mentalising (Senju *et al.*, 2009; Frith, 2012).

If you like interesting conversations, this is the floor to be. Mentalising supports rhetoric and ostensive communication, the hallmark of human social communication. By this I mean, that there is not simply an exchange of information between agents, but agents try and influence each other's minds. Inevitably there is a dark side, namely the aim to mislead the other individual, perhaps by concealing information, perhaps by implanting false beliefs. A video shows that clever birds can excel at this (Clayton, Dally & Emery, 2007). Let us glance also at some shining exhibits of beneficial processes that thrive on implicit mentalising: spontaneous helping (Buttelmann, Carpenter & Tomasello, 2009), automatic perspective taking (Samson *et al.*, 2010), and natural pedagogy (Csibra & Gergely, 2009). Spontaneous helping behaviour has been shown in human toddlers, because they recognise when others need help. It is relatively difficult to demonstrate in chimpanzees and other animals (Rekers, Haun & Tomasello, 2011).

What are the stimuli that trigger the implicit mentalising system? They are delivered through efficient dumb waiters from the ground floor and base-

ment. These are cues that suggest agency and are abundantly emitted by other agents. This includes eye gaze and facial expressions. Delays in mentalising performance by deaf and blind children suggest that trigger stimuli are multimodal and do not rely on one sense alone. On this floor the stimuli coming from other agents are sorted into intentional and accidental ones. Only intentional stimuli demand attention. Ostensive stimuli, such as calling your name or intently looking at you, instantly ignite brain regions that are part of the brain's mentalising system (Kampe, Frith & Frith, 2003).

How is the information processed that is signalled by the igniting stimuli? A model of an illuminated brain shows a highly connected network flashing when stimuli are delivered from the dumb waiters. We see lighting up medial prefrontal cortex, superior temporal sulcus at the temporal-parietal junction, parts of the basal temporal lobes, and amygdala regions (Frith & Frith, 2006). Studies have shown that in autism the connectivity in this network is much weaker (Castelli *et al.*, 2002; Kana *et al.*, 2009). There are some helpers provided by the lower floors. For instance, emotional contagion and the categorisation of another agent as a member of the in-group, facilitates attribution of mental states. This could be simply because the mental states are shared in this case, and it is not necessary to compute a different point of view.

While looking at the model you can see some sophisticated equations being displayed (Hampton, Bossaerts & O'Doherty, 2008; Yoshida, Dolan & Friston, 2008; Behrens, Hunt & Rushworth, 2009)). Maybe it is these equations that guide neuronal processes when they automatically compute causal explanations where causes are mental states, not physical states.

Top floor – Intelligence Services: explicit Theory of Mind

At last we come to the realm of Kahneman II. 'No children under 4 admitted' it says on the door. This floor is furnished in minimalist style, flooded with light and far removed from the boisterous bonhomie of the ground floor. Rational analytic thought pervades the atmosphere. Here we can find some tasteful plaques that document stages in the triumph of cultural evolution over biological evolution, awards for religion, for philosophy, for science, and many more. The firm of Kahneman II lives in relative isolation from the rest of the house but does intervene on occasion, although it is often too late and too little.

The main exhibit in one room is explicit mentalising. We see the props for classic false belief tasks, starting with Wimmer & Perner's (1983) Maxi paradigm. All test explicit mentalising ability: Maxi didn't know his mother had moved the chocolate and so he looks for it in a place where it no longer

is. This accomplishment is expected around the fifth year of life. The understanding of second order false beliefs, such as white lies and double bluff, does not lag far behind, but refinements are observed at later ages and adolescence represents a stage where changes in brain regions that support explicit mentalising show some radical changes (Dumontheil, Apperly & Blakemore 2010). Thus, the contribution of the medial prefrontal cortex in the network wanes in favour of the contribution of parts of the superior temporal cortex. Metacognition in the service of social communication is a work in progress. It can be argued that meta-cognition in the sense of 'thinking about thinking' is a logical consequence of explicit mentalising (Carruthers, 2009), giving rise to such notions as 'I think that I think and I believe that my thoughts are different from other people's thoughts' (Leslie, 1987).

Language is one of the most important tools for our self-awareness and acts as Chief Executive Officer in the firm of Kahneman II. It is indispensible for the explicit form of mentalising and drives its development. This is shown by the studies of mentalising ability over two generations of the Nicaraguan deaf community (Pyers & Senghas, 2009). Here, the first generation lacked any signs for mental states, and failed badly on standard false belief tasks. By the second generation, sign language had become established, and now everyone succeeded on the tasks.

Conscious thought and self-awareness depend on mechanisms that are not automatic, but require an effort to use, and that is why it is a slow system. When the wheels are turning, remarkably interesting processes take place. Education and learning has a vital role to keep them running. You can put on headphones to hear lectures on social etiquette; on how to make friends and influence people; on what it means to be responsible for your actions and what is involved in making a moral choice. The most impressive space is a gallery displaying architectural models of schools, universities, churches, law courts and museums. These models illustrate the enormous reach of Kahneman II over time to set up institutions that regulated group living and improve cooperation.

There are some secret rooms, labelled Security. Here a spymaster and a spin-doctor analyse the signals received from other houses for their truth value. They can also send strategically placed signals, for example, insincere flattery. A rogue's gallery shows cartoons telling of trickery, deception, persuasion and outright manipulation. There is an elaborate monitoring system dedicated to the business of reputation management (Tennie, Frith & Frith, 2010). Security services care for the maintenance of the house and decide on repairs and embellishments. Diplomacy rules here and this means that the system is flexible enough to forgive breeches of trust and solve social

dilemmas (Van Lange & Joireman, 2008). This system can override the rigid tit-for-tat mechanism operated on the ground floor.

I speculate that the flexible and even devious social processes in Kahneman II cannot be traced to innate social mechanisms. I find it unlikely that there are innate predispositions to sue rules politeness or to present diplomas to reward distinguished service. Instead these behaviours emerge at various stages of civilisation and culture. They can be lost, however useful they are in providing the oil to grease the machinery of complex social relationships in evolved societies.

There is another reason that I think no predispositions are needed to acquire explicit mentalising and all that follows from it. We know that in autism explicit mentalising can be acquired without any sign of implicit mentalising being present (Senju *et al.*, 2009). This separately acquired mentalising works best when off-line, for instance, using written communication with the cost of using a slow system. Of course, not all autistic individuals can become so accomplished, and this would be due to additional cognitive problems, which restrict basic information processing capacity. Individuals with impaired intellectual ability and without language struggle to acquire explicit mentalising and other tools that can be used for social manipulation. If there are cases where deficits in Kahneman II processes can be observed, I would not look for innate social mechanisms. Instead I would look for causes in lack of education, lack of cultural knowledge, and lack of general intellectual resources.

Attic – The Self

By a spiral staircase we can enter the attic and come to a tower. The surrounding walls are entirely made of mirrors and windows. The tower is the abode of the Self and my metaphor for the Self is a transparent balloon. It floats in the tower and is in constant need of being inflated. This is done by feeding its vanity (Sharot, Korn & Dolan, 2011). Obviously, the Self is selfish, but it can disguise selfishness by ostentatiously worn altruism. The Self is snobbish and identifies with conscious processes. But it has some inkling that there are many floors in the house. Its vocation is to rise above them. It can also ignore them, for instance when condemning prejudice even though an automatic in-group – out-group separation is still happening on the ground floor. The Self strives to be a free agent and it can occasionally impose its will and exert control over the rest of the house (Filevich, Kühn, & Haggard, 2012). However, the real power of control resides in many other switches on the lower floors that continuously turn on and off the complex machinery that sits there.

A sketch of the house is presented in Figure 3 (see p. 319). It is meant to serve as a mnemonic for the different mechanisms that may be resident in our social brain.

Social mechanisms in autism

Now that we have seen around the house, it is possible to think again about autism. In the following table I list social abilities that have been investigated in autism. I have ordered the studies in terms of their likelihood of being impaired in all or most individuals of the autism spectrum. My belief is that some of them will turn out to qualify as innate mechanisms, but this will depend on whether evidence can be obtained, first, from their clearly identifiable neural basis; second, from their specific dysfunction in well defined neuropsychological conditions; third, from their evolutionary origin In most cases, such evidence has not yet been obtained.

Table 1.

Social processes likely to be faulty in ASD

1. Eye gaze processing (Pelphrey *et al*., 2005; Grice *et al*., 2005; Kylliäinen *et al*., 2012)
2. Detecting biological motion (Kaiser & Pelphrey, 2012; Naeckaerts *et al*., 2012)
3. Mimicry/Rapport (Gallese *et al*., 2012; Scambler *et al*., 2007)
4. Self vs other distinction (Happé, 2010; Lombardo *et al*., 2010)
5. Mentalizing (Castelli *et al*., 2002; Kana *et al*., 2009; Yoshida *et al*., 2012)
    a. Perspective taking level 2 (Hamilton *et al*., 2009)
    b. Recognising social emotions (Shamey-Tsoory, 2008)
    c. Moral judgement (Moran *et al*., 2011; Gleichgerrcht *et al*., 2012)
    d. Reputation management (Izuma & Adolphs, 2011; Chevallier *et al*., 2012)

Social processes not likely to be faulty in ASD

1. Detecting agents (Johnson, 2003)
2. Identifying other's goals (Hamilton, 2009; Falck-Ytter, 2010)
3. Mirror neuron function (Hamilton, 2009; Leighton *et al*., 2008)
4. Attachment (Rutgers *et al*., 2004)
5. Salience of social stimuli (Fletcher-Watson, 2008; New *et al*., 2009)
6. Perspective taking, level 1 (Zwickel *et al*., 2011)
7. Cooperation, spontaneous helping (Colombi *et al*., 2009; Liebal *et al*., 2008)
8. Attributing social stereotypes of gender and race (Hirschfeld *et al*.)
9. Managing social hierarchies? (White *et al*., 2006)
10. Ingroup-outgroup formation? (Hirschfeld *et al*.)
11. Conformity? (Bowler & Worley, 1994)
12. Fairness/Inequity aversion? (Hill, Sally & Frith, 2004)

Social processes likely to be faulty, but not specific to ASD

1. Alexithymia (Silani *et al*., 2008; Bird *et al*., 2010)
2. Empathy (Singer *et al*., Jones *et al*., 2010; Bird *et al*., 2010)
3. Face processing (McPartland *et al*., 2011; Weigelt, Koldewyn, & Kanwisher, 2012)

As the Table shows in the case of autism the metaphorical house is perfectly sound in many ways. Sure, there are hits sustained by a number of separable social mechanisms. But there are also many putative social mechanisms that seem to be working well. Implicit mentalising is clearly the worst affected. However, it is interesting to speculate whether this mechanism is decomposable into more basic components and whether these too are compromised in autism. There are some promising computational models for mentalising (Hampton, Bossaerts & O'Doherty, 2008; Yoshida, Dolan & Friston, 2008; Behrens, Hunt & Rushworth, 2009), and relevant brain circuits have been pinpointed already.

The Table includes potential faults in top-down control mechanisms. For example, a number of researchers (e.g. Spengler, Bird & Brass, 2010; Cook & Bird, 2012; Grecucci *et al.*, 2012; Wang & Hamilton, 2012) have carried out studies of mimicry and argue that the problem here resides not in the mirror mechanism, but in a lack of top-down control. Chawarska, Makari & Shic (2012) take on the case of eye gaze and provide data that supports the hypothesis that lack of social attention is context dependent. Thus, they showed that autistic toddlers spend less time gazing at a face, relative to other children, only in situations where an adult makes a bid for their attention. Still even then, they look more at the adult compared to situations when there was no such bid. These types of argument point to new theories. These do not postulate problems in innate social mechanisms, but instead they point to problems in the control of these mechanisms. This is reminiscent of contemporary research on genetic diseases where the spotlight of attention has moved towards regions in the genome that switch on and off genes, rather than the genes themselves.

In the house metaphor I deliberately placed some highly active switches, for instance, between I-mode and We-mode. Perhaps a fault in this switch might explain what causes the characteristic egocentrism in autistic social interaction. To me autism has suggested an 'absent self', which is ironic, since I suggest that there is both too much and too little self (Frith, 2003; Frith, 2008). However, I could imagine that if the switch was stuck in the We-mode, there would be too little self, and if stuck in the I-mode, too much.

Two more points can be made through the Table. First, it is time to acknowledge that a remarkable amount of sociability can be present in autistic individuals, some of it still waiting to be revealed. Second, it is time to attend to possible overlap of pathologies in different disorders. Perhaps this is how we can understand that problems, such as lack of empathy for other people's feelings, or alexithymia, the inability to identify own feelings, are not restricted to autism but point to other types of pathology of the social brain.

Concluding remarks

The metaphor of the house serves to illustrate the complexity of the social mind/brain, up to a point. However, its main function may be to emphasise the need to search more deeply for the underlying neuro-cognitive mechanisms that are the secret drivers of our complex social world. What the metaphor makes easy to see is that there are likely to be many conditions of neuro-developmental origin, which are characterised by a variety of social deficits. The autism spectrum needs to be taken apart into distinct subgroups. Some of the subgroups may be characterised by failures in one or more of the social mechanisms themselves, while others may have problems only in the top-down control of these mechanisms.

While I believe that eventually a considerable number of innate social mechanisms will be found residing in Kahneman's System I, I also believe that most mechanisms in Kahneman's System II may not have an innate basis. So far no specific failure in this domain has been identified in autism over and above impairments in intellectual ability. For this reason I believe processes that are part of System II can usefully guide compensatory learning. Thus, autistic individuals can acquire explicit mentalising. The surprising picture that emerges from the mosaic of putative social mechanisms is that they seem rather independent. That is, a very basic mechanism, such as biological motion detection may be impaired, but this does not apparently affect the ready categorisation of in-groups and out-groups, nor the ability to show spontaneous empathy. It remains to be seen what kind of interdependence there is and what kind of compensatory learning can be achieved when biologically rooted impairments restrict social life.

References

[1] Abrahams B.S., Geschwind D.H. (2010). Connecting genes to brain in the autism spectrum disorders. *Arch Neurol*. 67(4):395-9. Review.

[2] Apperly, I. (2012). What is Theory of Mind? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, 65, 825-39.

[3] Apperly I.A., Butterfill S.A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychol Rev*. 116(4):953-70.

[4] Behrens T.E., Hunt L.T., Rushworth M.F. (2009). The computation of social behavior. *Science*. 324(5931):1160-4.

[5] Bernhard H., Fischbacher U., Fehr E. (2006). Parochial altruism in humans. *Nature*. 442(7105):912-5.

[6] Bird G., Press C., Richardson D.C. (2011). The role of alexithymia in reduced eye-fixation in Autism Spectrum Conditions. *41*. 41(11):1556-64.

[7] Bird, G., Catmur, C. Silani, G., Frith, C., & Frith, U. (2006). Attention does not modulate neural responses to social stimuli in autism spectrum disorders, *Neuroimage*, 31, 1614-24.

[8] Bird, G., Silani, G., Brindley, R., White, S. Frith, U., & Singer, T. (2010). Em-

pathic brain responses in insula are modulated by levels of alexithymia but not autism. *Brain*, 133(Pt 5),1515-25.

[9] Blair R.J. (2005). Responding to the emotions of others: dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition*, 14, 698-718.

[10] Bowler D.M. & Worley K. (1994). Susceptibility to social influence in adults with Asperger's syndrome: a research note. *J Child Psychol Psychiatry*. 35(4): 689-97.

[11] Boyd R., Richerson P.J., Henrich J. (2011). The cultural niche: why social learning is essential for human adaptation. *Proc. Natl. Acad. Sci. U.S.A.* 108 Suppl 2:10918-25.

[12] Boyd R., Richerson P.J. (2009). Culture and the evolution of human cooperation. *Philos Trans R Soc Lond B Biol Sci*. 364(1533):3281-8.

[13] Bugnyar T. (2011). Knower-guesser differentiation in ravens: others' viewpoints matter. *Proc Biol Sci*. 278(1705): 634-40.

[14] Burgess, N. & O'Keefe, J. (2011). Models of place and grid cell firing and theta rhythmicity. *Curr Opin Neurobiol*. 21(5):734-44.

[15] Buttelmann D., Carpenter M., Tomasello M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112(2);337-42.

[16] Carruthers P. (2009). How we know our own minds: the relationship between mindreading and metacognition. *Behav Brain Sci*. 32(2):121-38; discussion 138-82.

[17] Carter C.S., Grippo A.J., Pournajafi-Nazarloo H., Ruscio M.G., Porges S.W. (2008). Oxytocin, vasopressin and sociality. *Prog Brain Res*. 170:331-6.

[18] Castelli, F., Frith, C.D., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, 125, 1839-1849.

[19] Chawarska, K. Makari S. & Shic F. (2012). Context modulates attention to social scenes in toddlers with autism. *J Child Psychol Psychiatry*. 53(8):903-13.

[20] Chevallier C., Kohls G., Troiani V., Brodkin E.S., Schultz R.T. (2012). The social motivation theory of autism. *Trends Cogn Sci*. 16(4):231-9.

[21] Chevallier, C., Molesworth, C. & Happé. F. (2012). Diminished social motivation negatively impacts reputation management: Autism Spectrum Disorders as a case in point. *PLoS ONE*, 7 (2012), p. e3110

[22] Christakis N.A. & Fowler J.H. (2012). Social contagion theory: examining dynamic social networks and human behavior. *Stat Med*. 2012 E-Pub.

[23] Cialdini, R.B., & Goldstein, N.J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55, 591-621.

[24] Clayton N.S., Dally J.M., Emery N.J. (2007). Social cognition by food-caching corvids. The western scrub-jay as a natural psychologist. *Philos Trans R Soc Lond B Biol Sci*. 362(1480):507-22.

[25] Colombi C., Liebal K., Tomasello M., Young G., Warneken F., Rogers S.J. (2009). Examining correlates of cooperation in autism: Imitation, joint attention, and understanding intentions. *Autism*. 13(2):143-63.

[26] Cook J.L., Bird G. (2012). Atypical social modulation of imitation in autism spectrum conditions. *J Autism Dev Disord*. 42(6):1045-51.

[27] Cook & Mineka S. (1989). Observational conditioning of fear to fear-relevant versus fear-irrelevant stimuli in rhesus monkeys. *J Abnorm Psychol*. 98(4):448-59.

[28] Csibra G., Gergely G. (2009). Natural pedagogy. *Trends Cogn Sci*. 13(4):148-53.

[29] Delgado M.R. (2007). Reward-related responses in the human striatum. *Ann NY Acad Sci*. 1104:70-88.
[30] Dumontheil I., Apperly I.A., Blakemore S.J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Dev Sci*. 13(2):331-8.
[31] Emery N.J., Clayton N.S. (2009). Comparative social cognition. *Annu Rev Psychol*. 60:87-113.
[32] Falck-Ytter, T. (2010). Young children with autism spectrum disorder use predictive eye movements in action observation. *Biol Lett*. 6(3):375-8.
[33] Fehr E., Bernhard H., Rockenbach B. (2008). Egalitarianism in young children. *Nature*. 454(7208):1079-83.
[34] Filevich, E., Kühn, S. & Haggard, P. (2012). Intentional inhibition in human action: the power of 'no'. *Neurosci Biobehav Rev*. 36(4):1107-18.
[35] Fletcher-Watson S., Leekam S.R., Findlay J.M., Stanton E.C. (2008). Brief report: young adults with autism spectrum disorder show normal attention to eye-gaze information-evidence from a new change blindness paradigm. *J Autism Dev Disord*. 38(9):1785-90.
[36] Frith, C.D., Frith, U. (2006). The neural basis of mentalizing. *Neuron*, 50(4) 531-4.
[37] Frith C.D. & Frith U. (2008). Implicit and explicit processes in social cognition. *Neuron*. 2008 Nov 6;60(3):503-10.
[38] Frith, C.D. & Frith, U. (2012). Mechanisms of Social Cognition. *Annual Review of Psychology*, 63, 287-313
[39] Frith, U. (2004). Emmanuel Miller lecture: Confusions and controversies about Asperger syndrome, *Journal of Child Psychology and Psychiatry*, 45, 672-686.
[40] Frith, U. (2008). *A Very Short Introduction to Autism*. Oxford: Oxford University Press
[41] Frith, U. (2012). Why we need cognitive explanations of autism. *Quarterly Journal of Experimental Psychology* [August 21 E-pub].
[42] Frith, U. (2001). Mindblindness and the brain in Autism. *Neuron*, 32, 969-979.
[43] Gallese V, Rochat MJ, Becchio C. (2012). The mirror mechanism and its potential role in autism spectrum disorder. *Dev Med Child Neurol*. E-pub.
[44] Gavrilets S. (2012). On the evolutionary origins of the egalitarian syndrome. *Proc. Natl. Acad. Sci. U.S.A.* 109(35):14069-74.
[45] Geschwind D.H. (2009). Advances in autism. *Annu Rev Med*. 60:367-80.
[46] Gleichgerrcht E., Torralva T., Rattazzi A., Marenco V., Roca M., Manes F. (2012). Selective impairment of cognitive empathy for moral judgment in adults with high functioning autism. *Soc Cogn Affect Neurosci*. E-Pub.
[47] Grecucci A., Brambilla P., Siugzdaite R., Londero D., Fabbro F., Rumiati R.I. (2012). Emotional Resonance Deficits in Autistic Children. *J Autism Dev Disord*. E-Pub.
[48] Grice S.J., Halit H., Farroni T., Baron-Cohen S., Bolton P., Johnson M.H. (2005). Neural correlates of eye-gaze detection in young children with autism. *Cortex*. 41(3):342-53.
[49] Gotts S.J., Simmons W.K., Milbury L.A., Wallace G.L., Cox R.W., Martin A. (2012). Fractionation of social brain circuits in autism spectrum disorders. *Brain*. 135(Pt 9):2711-25.
[50] Hamilton A.F. (2009). Goals, intentions and mental states: challenges for theories of autism. *J Child Psychol Psychiatry*. 50(8):881-92.
[51] Hamilton, A.F., Brindley, R., & Frith, U. (2009). Visual perspective taking impairment in children with autistic spectrum disorder. *Cognition*, 113, 37-44.
Hampton A.N., Bossaerts P., O'Doherty J.P.

(2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc. Natl. Acad. Sci. U.S.A.* 105(18):6741-6.

Happé, F. (200) Theory of mind and the self. *Ann NY Acad Sci.* 1001:134-44.

Haruno M., Frith C.D. (2010). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nat Neurosci.* 13(2):160-1.

Haun, D., van Leeuwen, E. & Edelson, M. (2012). Majority influence in children and other animals. *Dev Cog Neurosci.* E-pub.

Hermans E.J., van Wingen G., Bos P.A., Putman P., van Honk J. (2009). Reduced spontaneous facial mimicry in women with autistic traits. *Biol Psychol.* 80(3):348-53.

Heyes C. (2011). Automatic imitation. *Psychol Bull.* 201137(3):463-83.

Heyes C. (2012). Grist and mills: on the cultural origins of cultural learning. *Philos Trans R Soc Lond B Biol Sci.* 367(1599):2181-91.

Hill, E.L., Berthoz, S., & Frith, U. (2004). Brief Report: Cognitive processing of own emotions in individuals with autistic spectrum disorder and their relatives. *Journal of Autism and Developmental Disorders*, 34, 229-235.

Hill, E.L., Sally, D. and Frith, U. (2004). Does mentalising ability influence cooperative decision making in a social dilemma? Introspective evidence from a study of adults with autism spectrum disorder. *Journal of Consciousness Studies*, 11, 1-18.

Hirschfeld L., Bartmess E., White S. & Frith U. (2007). Can autistic children predict behavior by social stereotypes? *Current Biology*, 17(12):R451-2.

Izuma K., Matsumoto K., Camerer C.F., Adolphs R. (2011). Insensitivity to social reputation in autism. *Proc. Natl. Acad. Sci. U.S.A.*, 108. 17302-17307.

Johnson, S.C. (2003). Detecting agents. *Philos Trans R Soc Lond B Biol Sci.* 358(1431):549-59.

Kahneman, D. (2011). *Thinking, Fast and Slow.* N.Y.: Straus and Giroux.

Kaiser M.D., & Pelphrey K.A. (2012). Disrupted action perception in autism: behavioral evidence, neuroendophenotypes, and diagnostic utility. *Dev Cogn Neurosci.* 2(1):25-35.

Kampe, K., Frith, C.D., Dolan, R.J., Frith, U. (2001). Attraction and gaze – the reward value of social stimuli. *Nature*, 413, 589.

Kana R.K., Keller T.A., Cherkassky V.L., Minshew N.J., Just M.A. (2009). Atypical frontal-posterior synchronization of Theory of Mind regions in autism during mental state attribution. *Soc Neurosci.* 4(2):135-52.

Kanwisher N., McDermott J., Chun M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci* 17:4302-11.

Kilner J.M., Friston K.J., Frith C.D. (2007). The mirror-neuron system: a Bayesian perspective. *Neuroreport.* 18(6):619-23.

Knill D.C. & Pouget A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27(12):712-9.

Knoblich G., Sebanz N. (2008). Evolving intentions for social interaction: from entrainment to joint action. *Philos Trans R Soc Lond B Biol Sci.* 363(1499):2021-31.

Kovács Á.M., Téglás E., Endress A.D. (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science*, 330(6012):1830-4.

Krach, S., Paulus, F.M., Bodden, M. & Kircher, T. (2010). The rewarding nature of social interactions. *Front Behav Neurosci.* 4:22.

Kylliäinen A., Wallace S., Coutanche M.N., Leppänen J.M., Cusack J., Bailey A.J., Hietanen J.K. (2012). Affective-motivational brain responses to direct gaze

in children with autism spectrum disorder. *J Child Psychol Psychiatry*. 53(7):790-7.

Leighton J., Bird G., Charman T., Heyes C. (2008). Weak imitative performance is not due to a functional 'mirroring' deficit in adults with AutismSpectrum Disorders. *Neuropsychologia*. 46(4): 1041-9.

Leslie, A.M. (1987) Pretense and representation. The origin of "theory of mind". *Psychological Review*, 94, 412-426.

Liebal K., Colombi C., Rogers S.J., Warneken F., Tomasello M. (2007). Helping and cooperation in children with autism. *J Autism Dev Disord*. 38(2):224-38.

Lombardo, M.V., Chakrabarti, B., Bullmore, E.T., Sadek, S.A., Pasco, G., Wheelwright, S.J., Suckling, J., MRC AIMS Consortium, Baron-Cohen, S. (2010). Atypical neural self-representation in autism. *Brain*, 133, 611-24.

McCall C., Singer T. (2012). The animal and human neuroendocrinology of social cognition, motivation and behavior. *Nat Neurosci*. 15(5):681-8.

McPartland J.C., Webb S.J., Keehn B., Dawson G. (2011). Patterns of visual attention to faces and objects in autism spectrum disorder. *J Autism Dev Disord*. 41(2):148-57.

Minio-Paluello I., Baron-Cohen S., Avenanti A., Walsh V., Aglioti S.M. (2009). Absence of embodied empathy during pain observation in Asperger syndrome. *Biol Psychiatry*. 65(1):55-62.

Moran J.M., Young L.L., Saxe R., Lee S.M., O'Young D., Mavros P.L., Gabrieli J.D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proc Natl Acad Sci U S A*. 108(7):2688-92.

Morishima, Y. Schunk, D., Bruhin, A., Ruff, C.C. & Fehr, E. (2012). Linking brain structure and activation in the temporoparietal junction to explain the neurobiology of human altruism. *Neuron*, E-pub.

Morton, J. & Frith, U. (1995). Causal modeling: Structural approaches to developmental psychopathology. In D. Cicchetti, & D. Cohen (Eds.), *Developmental Psychopathology* (pp. 357-90). New York, NY: Wiley.

Nackaerts E., Wagemans J., Helsen W., Swinnen S.P., Wenderoth N., Alaerts K. (2012). Recognizing biological motion and emotions from point-light displays in autism spectrum disorders. *PLoS One*. 7(9).

New J.J., Schultz R.T., Wolf J., Niehaus J.L., Klin A., German T.C., Scholl B.J. (2009). The scope of social attention deficits in autism: prioritized orienting to people and animals in static natural scenes. *Neuropsychologia*. 48(1):51-9.

Nowak M.A., Sigmund K. (1998). The dynamics of indirect reciprocity. *J Theor Biol*. 21;194(4):561-74.

Onishi, H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 8, 308(5719), 255-8.

Over, H. & Carpenter, M. (2008). Priming third-party ostracisms increases affiliative imitation in children. *Dev Sci*. 12(3) F1-F8.

Pelphrey, K.A., Morris, J.P., & McCarthy, G. (2005). Neural basis of eye gaze processing deficits in autism. *Brain*, 128(Pt 5):1038-48.

Pelphrey K.A., Shultz S., Hudac C.M., Vander Wyk B.C. (2011). Research review: Constraining heterogeneity: the social brain and its development in autism spectrum disorder. *J Child Psychol Psychiatry*. 52(6):631-44.

Petticrew M., Davey Smith G. (2012). The monkey puzzle: a systematic review of studies of stress, social hierarchies, and heart disease in monkeys. PLoS One. 2012;7(3)

Piazza M., Pinel P., Le Bihan D., Dehaene S. (2007) A magnitude code common

to numerosities and number symbols in human intraparietal cortex. *Neuron*, 18;53(2):293-305.

Press C., Richardson D., Bird G. (2010). Intact imitation of emotional facial actions in autism spectrum conditions. *Neuropsychologia*. 48(11):3291-7. Epub 2010 Jul 16.

Pyers J.E., Senghas A. (2009). Language promotes false-belief understanding: evidence from learners of a new sign language. *Psychol Sci*. 20(7):805-12.

Raafat R.M., Chater N., Frith C. (2009). Herding in humans. *Trends Cogn Sci*. 13(10):420-8.

Raihani N.J., Bshary R. (2011). Resolving the iterated prisoner's dilemma: theory and reality. *J Evol Biol*. 24(8):1628-39.

Raihani N.J., McAuliffe K. (2012). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biol Lett*. 8(5):802-4.

Rekers Y., Haun D.B., Tomasello M. (2011). Children, but not chimpanzees, prefer to collaborate. *Curr Biol*. 21(20):1756-8.

Rendell L., Boyd R., Cownden D., Enquist M., Eriksson K., Feldman M.W., Fogarty L., Ghirlanda S., Lillicrap T., Laland K.N. (2010). Why copy others? Insights from the social learning strategies tournament. *Science*. 328(5975):208-13.

Rutgers A.H., Bakermans-Kranenburg M.J., van Ijzendoorn M.H., van Berckelaer-Onnes I.A. (2004). Autism and attachment: a meta-analytic review. *J. Child Psychol. Psychiatry*, 45, 1123-1134

Samson D., Apperly I.A., Braithwaite J.J., Andrews B.J., Bodley Scott S.E. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people see. *J Exp Psychol Hum Percept Perform*. 36(5):1255-66.

Scambler D.J., Hepburn S., Rutherford M.D., Wehner E.A., Rogers S.J. (2007). Emotional responsivity in children with autism, children with other developmental disabilities, and children with typical development. *J Autism Dev Disord*. 37(3):553-63.

Shamay-Tsoori, S.G. (2008). Recognition of 'fortune of others' emotions in Asperger syndrome and high functioning autism. *J Autism Dev Disord*. 38(8):1451-61.

Schneider D., Lam R., Bayliss A.P., Dux P.E. (2012). Cognitive Load Disrupts Implicit Theory-of-Mind Processing. *Psychol Sci*, 23(8):842-7.

Senju, A., Southgate, V., White, S. and Frith, U. (2009). Mindblind eyes: an absence of spontaneous Theory of Mind in Asperger Syndrome, *Science*, 325(5942):883-5.

Sharot, T., Korn, C.W. & Dolan, R. (2011). How unrealistic optimism is maintained in the face of reality. *Nat Neurosci*. 14(11):1475-9.

Sheridan C. (2011). Gene therapy finds its niche. *Nat Biotechnol*. 29(2):121-8.

Shkurko (2012). Is social categorization based on relational ingroup/outgroup opposition? A meta-analysis. *Soc Cogn Affect Neurosci*. 2012 Jul 30.

Silani, G., Bird, G., Brindley, R., Singer, T., Frith, C., & Frith, U. (2008). Levels of emotional awareness and autism: an fMRI study. *Social Neuroscience*, 3, 97-112.

Singer T., Lamm C. (2009). The social neuroscience of empathy. *Ann. N.Y. Acad. Sci.*, 1156:81-96.

Tabibnia, G. & Lieberman, M.D. (2007). Fairness and cooperation are rewarding. *Ann. N.Y. Acad. Sci.*, 1118, 90-101.

Tennie C., Frith U., Frith C.D. (2010). Reputation management in the age of the world-wide web. *Trends Cogn Sci*, 14, 482-8.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge Mass: MIT Press.

Tomasello M, Vaish A. (2012). Origins of Human Cooperation and Morality. *Annu Rev Psychol*. Jul 12.

Tuomela, R. (2007). *The philosophy of so-*

ciality. Oxford Scholarship Online. September 2007.

van Baaren, R.B., Holland, R.W., Kawakami, K., & van Knippenberg, A. (2004). Mimicry and prosocial behavior. *Psychological Science*, 15(1), 71-74.

van Lange, P. & Joireman, J.A. (2008). How we can promote behaviour that serves all of us in the future. *Social Issues and Policy Review*, 2(1):127-57.

Wang Y. & Hamilton, A. (2012). Social top-down response modulation (STORM): a model of the control of mimicry in social interaction. *Front Hum Neurosci*. 6:153. Epub.

Warneken F., Tomasello M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*. 311(5765):1301-3.

Weigelt S., Koldewyn K., Kanwisher N., (2012). Face identity recognition in autism spectrum disorders: a review of behavioral studies. *Neurosci Biobehav Rev.* 2012 Mar;36(3):1060-84.

White S., Hill E., Winston J., Frith U. (2006). An islet of social ability in Asperger Syndrome: judging social attributes from faces. *Brain Cogn*. 61(1):69-77.

Williams J.H. (2008). Self-other relations in social development and autism: multiple roles for mirror neurons and other brain bases. *Autism Res.* 1(2):73-90.

Yoshida W., Dolan R.J., Friston K.J. (2008). Game theory of mind. *PLoS Comput Biol*. 4(12).

Yoshida W., Dziobek I., Kliemann D., Heekeren H.R., Friston K.J., Dolan R.J. (2010). Cooperation and heterogeneity of the autistic mind. *J Neurosci*. 30(26):8815-8.

Zink C.F., Meyer-Lindenberg A. (2012). Human neuroimaging of oxytocin and vasopressin in social cognition. *Horm Behav*. 61(3):400-9. 3

Zink C.F., Tong Y., Chen Q., Bassett D.S., Stein J.L., Meyer-Lindenberg A. (2008). Know your place: neural processing of social hierarchy in humans. *Neuron*. 58(2):273-83.

Zwickel J., White S.J., Coniston D., Senju A., Frith U. (2011). Exploring the building blocks of social cognition: spontaneous agency perception and visual perspective taking in autism. *Soc Cogn Affect Neurosci*. 6(5):564-71.

Natural Cooperation

Martin A. Nowak[1]

Evolution is a process which organizes the living world. Loosely speaking we talk about the evolution of genes, genomes, cells, organisms, species, but the only entity that really evolves are populations. Populations of reproducing individuals instantiate the evolutionary process. Individuals carry information which they pass on during reproduction. This process of reproduction is not perfectly accurate but subject to variation. Thereby new mutants are generated. If different mutants have different reproductive rates, natural selection comes into play. Natural selection chooses among the variants that are generated by mutation. In the classical formulation, mutation and selection are the two fundamental components of the evolutionary process.

In recent years I have proposed that cooperation can be seen a third fundamental component of evolution. Cooperation means that two individuals, who are competitors in the process of natural selection, help one another. Without cooperation there is no construction. Cooperation is present at the origin of life, when nucleotide sequences help each other to reproduce within protocells. Cooperation is involved when individual cells stay together to form the first multi-cellular organism. Cancer is a breakdown of cooperation among the cells of a multi-cellular organism. Cooperation is needed for the emergence of the superorganism of insect societies, which represents a distinct form of biological organization. Cooperation is crucial for the evolution of human society and human language.

In the absence of a specific mechanism, natural selection opposes cooperation. In any well-mixed population defectors have a higher payoff than unconditional cooperators. Therefore natural selection needs help to favor cooperation over defection. Thousands of scientific papers have been written on this topic. All suggestions so far can be categorized into five mechanisms, which I will now discuss. A mechanism for the evolution of cooperation is an interaction structure, specifying how the individuals of a population interact to accumulate payoff and compete for reproduction.

1) Direct reciprocity arises if there are repeated encounters between the same two individuals, who use conditional strategies that depend on previ-

[1] Program for Evolutionary Dynamics, Departments of Mathematics, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA.

ous outcomes. Direct reciprocity is based on the concept of repeated games and embodies the simply idea: I help you and you help me. Successful strategies of direct reciprocity include generous tit-for-tat and win-stay, lose-shift. Generous tit-for-tat starts with cooperation, it always cooperates if the other person has cooperated, and it sometimes cooperates even if the other person has defected. Win-stay, lose-shift repeats its current move whenever it does well, but changes to the opposite move whenever it does badly.

2) Indirect reciprocity operates if there are repeated encounters in a population of individuals. Some encounters are observed by others. Information about those encounters spreads through communication. Individuals can adopt conditional strategies that base their decision on the reputation of the recipient. My behavior towards you depends on what you have done to me and to others. The key aspect of indirect reciprocity is reputation. Cooperation is costly but leads to the reputation of a helpful individual who might receive help from others. A strategy for indirect reciprocity consists of a social norm and an action rule. The social norm specifies how to evaluate interactions between individuals. The action rule specifies whether or not to cooperate given the information about the other individual. Indirect reciprocity can lead to cooperation, if the probability to know someone's reputation is sufficiently high.

3) Spatial selection can lead to the evolution of cooperation without strategic complexity. Behaviors need not be conditional on previous outcomes. Cooperators form clusters which prevail even if they are surrounded by defectors. The fundamental idea is that neighbors help each other. More generally, population structure affects the outcome of the evolutionary process, and some population structures can lead to the evolution of cooperation. The population structure can be static or dynamic. It can represent geographic distribution or social networks. For example, evolutionary graph theory studies evolutionary dynamics on static graphs, while evolutionary set theory describes individuals moving between sets thereby changing the interaction structure as part of the evolutionary process.

4) Multi-level selection operates if there is competition between individuals in a group and competition between groups. It is possible that defectors win within groups, but that groups of cooperators outcompete groups of defectors. Overall this process can result in the selection of cooperators. Darwin wrote in 1871: 'There can be no doubt that a tribe including many members who ... were always ready to give aid to each other and to sacrifice themselves for the common good, would be victorious over other tribes; and this would be natural selection'.

5) Kin selection can be seen as a mechanism for the evolution of cooperation if properly formulated. In my opinion, kin selection operates if there

is conditional behavior based on kin recognition. An individual recognizes kin and behaves accordingly. As JBS Haldane said 'I will jump into the river to save two brothers or eight cousins'. Unfortunately much of the current kin selection literature does not adhere to this simple definition. Instead kin selection is often linked to the concept of inclusive fitness, which is a particular method to account for fitness effects. Inclusive fitness works in special cases, but is mistakenly presented as a general concept. When studying social evolution it is best not to rely on inclusive fitness. Once fitness is calculated every aspect of relatedness is included. Kin selection requires a mathematical formulation which is not limited by inclusive fitness.

These are five mechanisms for the evolution of cooperation. There may be others. But so far all suggestions fall within these mechanisms. Often two or more mechanisms operate simultaneously, which can lead to synergistic effects. When discussing human behavior it is important to note that much of the current theory examines actions and responses to actions, but not motivation. In my opinion, human altruism can only be understood by examining the underlying motivation. An action is truly altruistic if motivated by love for the other person. This is difficult to study, but an important direction for future research.

Evolution is based on the three fundamental principles: mutation, selection and cooperation. Evolution is a search process. Every search process requires a search space, a space of limited possibilities that is being explored. Much discussion in evolutionary biology is about the search process. The molecular components of biological organisms (DNA, RNA, proteins etc) point toward the nature of the underlying search space for genetic evolution, but how exactly this search space is generated by the laws of physics and chemistry is elusive at present.

References and further reading

Alexander, R.D. 1987. *The Biology of Moral Systems*, New York: Aldine de Gruyter.

Axelrod, R., 1984. *The evolution of cooperation*. Basic Books New York.

Dreber, A., Rand, D., Fudenberg, D., Nowak, M.A., 2008. Winners don't punish. *Nature* 452, 348-352.

Fudenberg, D., Maskin, E., 1990. Evolution and Cooperation in Noisy Repeated Games. *American Economic Review* 80, 274-279.

Gadagkar, R., 2010. Sociobiology in turmoil again. *Current Science* 99, 1036-1041.

Hamilton, W.D., 1964. The genetical evolution of social behaviour, I and II. *J. Theor. Biol.* 7, 1-52.

Hardin, G., 1968. The tragedy of the commons. *Science* 3859, 1243-1248.

Hauert, C., De Monte, S., Hofbauer, J., Sigmund, K., 2002. Volunteering as red queen mechanism for cooperation in public goods games. *Science* 296, 1129-1132.

Hofbauer, J., Sigmund, K., 1998. *Evolutionary games and population dynamics*. Cambridge

University Press, Cambridge, UK.
Kerr, B., Godfrey-Smith, P., 2002. Individualist and multi-level perspectives on selection in structured populations. *Biol. Philos.* 17, 477-517.
Lieberman, E., Hauert, C., Nowak, M.A., 2005. Evolutionary dynamics on graphs. *Nature* 433, 312-316.
May, R.M. 1987. More evolution of cooperation. *Nature* 327: 15-17.
Maynard Smith, J., 1982. *Evolution and the theory of games*. Cambridge University Press, Cambridge, UK.
Milinski, M., 1987. Tit for tat in sticklebacks and the evolution of cooperation. *Nature* 325, 433-435.
Nowak, M.A., May, R.M., 1992. Evolutionary games and spatial chaos. *Nature* 359, 826-829.
Nowak, M.A., Sigmund, K. 1992. Tit for tat in heterogeneous populations. *Nature* 355, 250-253.
Nowak, M.A., Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. *Nature* 393, 573-577.
Nowak, M.A., 2006. Five rules for the evolution of cooperation. *Science* 314, 1560-1563.
Nowak, M.A., Tarnita, C.E., Wilson, E.O., 2010. The evolution of eusociality. *Nature* 466, 1057-1062.
Ohtsuki, H., Hauert, C., Lieberman, E., Nowak, M.A., 2006. A simple rule for the evolution of cooperation on graphs and social networks. *Nature* 441, 502-505.
Ohtsuki, H., Iwasa, Y., 2006. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* 239, 435-444.
Ostrom, E., 1990. *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press, Cambridge, UK.
Perc, M., 2009. Evolution of cooperation on scale-free networks subject to error and attack. *New J. Phys.* 11, 033027.
Rand, D.G., Dreber, A., Ellingson, T., Fudenberg, D., Nowak, M.A., 2009. Positive interactions promote public cooperation. *Science* 325, 1272-1275.
Rapoport, A., Chammah, A.M., 1965. *Prisoner's dilemma*. University of Michigan Press, Ann Arbor MI.
Rockenbach, B., Milinski, M., 2006. The efficient interaction of indirect reciprocity and costly punishment. *Nature* 444, 718-723.
Seinen, I., Schram, A., 2005. Social status and group norms: indirect reciprocity in a repeated helping experiment. *European Econ. Rev.* 50, 581-602.
Sigmund, K., 2010. *The calculus of selfishness*. Princeton University Press, Princeton, NJ.
Sober, E., Wilson, D.S., 1999. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press
Sugden, R. 1986. *The Economics of Rights, Cooperation and Welfare*. Oxford: Basil Blackwell.
Szabó, G., Fáth, G., 2007. Evolutionary games on graphs. *Phys. Rep.* 446, 97-216.
Szathmáry, E., Demeter, L., 1987. Group selection of early replicators and the origin of life. *J. Theor. Biol.* 128, 463-486.
Tarnita, C.E., Antal, T., Ohtsuki, H., Nowak, M.A., 2009. Evolutionary dynamics in set structured populations. *P. Natl. Acad. Sci. USA* 106, 8601-8604.
Traulsen, A., Nowak, M.A., 2006. Evolution of cooperation by multi-level selection. *P. Natl. Acad. Sci. USA* 103, 10952-10955
Trivers, R.L., 1971. The evolution of reciprocal altruism. *Q. Rev. Biol.* 46, 35-57.
Wade, M.J., 1977. An experimental study of group selection. *Evolution* 31, 134-153.
Wedekind, C., Milinski, M., 2000. Cooperation through image scoring in humans. *Science* 288, 850-852.
Weibull, J.W., 1995. *Evolutionary game theory*. MIT Press, Cambridge, MA.
Wilson, D.S., 1975. *A theory of group selection*. Proc. Natl Acad. Sci. USA 72, 143-146.

Spiritualité de l'âme

Georges Card. Cottier, O.P.

Introduction

1. Une double distinction commande l'approche de la vérité dans la considération que je vous propose.

En premier lieu, il existe deux sources de notre connaissance de la vérité: la raison naturelle et la Révélation divine. L'objet de la théologie est de scruter rationnellement le contenu de la Révélation. Le savoir acquis par les sciences et par la philosophie procède de la raison s'appuyant sur ses propres ressources.

De droit, entre ces deux sources, il ne peut y avoir contradiction. Ayant toutes deux leur origine ultime en Dieu, elles sont complémentaires. L'intelligence de la foi (ou théologie) présuppose la reconnaissance de l'autonomie de la raison dans son ordre propre.

Dans ce qui suit, notre réflexion se situe au niveau de la rationalité philosophique.

En second lieu, l'activité de la raison se diversifie en fonction de la nature des objets qu'elle interroge, du type de questions qu'elle pose à la réalité. Chacune de ces différentes approches requiert sa propre méthode. C'est pourquoi l'œuvre de la raison se diversifie en une pluralité de savoir articulés entre eux. C'est là une donnée épistémologique de base. Le scientisme, qui est une tentation répandue, pousse à réduire le champ de la raison et du rationnel au seul domaine des sciences expérimentales et à leur formalisation mathématique.

Ce que nous proposons est de nature philosophique, se rattachant plus précisément à l'anthropologie philosophique.

2. Ajoutons que si la philosophie jouit d'une légitime autonomie par rapport à la Révélation, elle a, comme le montre l'histoire, grandement bénéficié de son contact avec elle. Elle lui doit une impulsion décisive pour l'approfondissement de sa connaissance de l'homme et de sa dignité. C'est ainsi que le thème biblique de l'homme *créé à l'image de Dieu* est à la racine de la conception de l'individu humain comme personne.

Substance raisonnable, l'homme est un sujet ayant la maîtrise de ses actes et la capacité de décider. Agissant par soi, il est appelé à rendre raison de ses actes; il en est responsable, ce qui est la marque de sa liberté.

Par nature il est apte à nouer des relations avec autrui; dans les autres, il reconnaît des sujets qui lui sont égaux. La société humaine est une société de personnes.

Ayant ses racines dans la transcendance, il est capable de rentrer en relation de *je* et *tu* avec Dieu.

Au titre de personne, qui a une portée ontologique, il ne peut être une simple émanation des énergies immanentes de la matière. Son âme spirituelle, principe de son être de personne, est directement créée par Dieu. C'est ce dernier point que nous essayerons d'élucider.

L'âme

3. Le concept d'âme (*psyché, anima*) vient de la tradition philosophique grecque. Il est au cœur de la pensée de Platon. Il a été repris par Aristote qui en a précisé la signification.

La réflexion philosophique porte sur les faits premiers de l'existence. Ainsi notre regard sur la réalité, celle qui est autour de nous et celle qui est en nous, est frappé par le fait de la mobilité et de l'instabilité des choses. Aristote s'est ainsi interrogé sur le mouvement, au sens large de mutation, et sur ses causes. Cette réflexion le conduit à discerner dans les êtres que nous connaissons, un double principe: la matière et la forme. La matière est le substrat, le principe d'indétermination apte à recevoir les déterminations qu'elle ne peut se donner à elle-même. La forme est le principe qui apporte cette détermination; elle donne à un être sa structure, son organisation interne, et son actualité. Tous les êtres de la nature sont composés de matière et de forme.

L'approfondissement métaphysique des notions de matière et de forme conduira ensuite le philosophe à reconnaître qu'elles se rattachent à une division de l'être de portée universelle, la division entre *puissance*, *potentialité* et *acte*.

La distinction des êtres n'est pas uniquement numérique et quantitative. Elle est également spécifique. C'est là un second fait premier sur lequel portera la réflexion du philosophe. Les êtres ou groupes d'êtres, ont leur propre identité, leurs caractéristiques, qualités et activités, distinctives. C'est la forme qui donne à un être son identité spécifique.

Dans l'échelle des êtres, certains possèdent des propriétés et des activités autres que celles qui sont communes à tous les êtres de la nature: les êtres *vivants*.

Aristote appelle âme la forme qui fait du vivant un être vivant. Cette forme qu'est l'âme donne au vivant son unité, son organisation interne, ses capacités d'agir et ses organes. Grâce à ses organes, le vivant agit comme sujet d'action un.

Dans le *traité de l'âme* (*peri psychè*), livre II, 412 a 27, Aristote donne la définition suivante de l'âme: l'âme est l'acte (*entéléchia*) premier d'un corps naturel, ayant la vie en puissance, c'est-à-dire d'un corps organique.

La puissance, la potentialité elle-même comporte des degrés allant de la pure indétermination (*materia prima*) à la pleine actualité. L'acte premier désigne l'acte qui constitue la substance.

Mais celle-ci est en disposition d'acquérir des perfections ultérieures, par l'actualisation des puissances ou facultés ordonnées aux divers types d'activité.

Selon l'ordre de perfection de ces activités, Aristote est conduit à reconnaître trois ordres de vivants: l'ordre végétal, l'ordre animal, l'ordre intellectuel (c'est-à-dire celui de l'*animal raisonnable*). A ces trois ordres correspondent trois types d'âmes: l'âme végétative, l'âme sensitive, l'âme intellectuelle.

L'âme des vivants supérieurs intègre les activités des vivants inférieurs. C'est ainsi que l'âme intellectuelle est aussi principe des activités végétatives et sensitives. L'unité du sujet requiert l'unité de l'âme. Il importe de le souligner: l'âme, au sens aristotélicien se définit en fonction du corps (le corps vivant).

La doctrine de l'âme chez Thomas d'Aquin se présente comme un développement des vues d'Aristote.

L'âme a pour ainsi dire disparu de la réflexion philosophique contemporaine. Un des facteurs de cette éclipse est la conception cartésienne de l'âme. Descartes réserve le terme uniquement pour désigner l'âme spirituelle (*Moi, c'est-à-dire mon âme*). Sa distinction du corps, dont elle n'est pas la forme, s'entend dans le sens du dualisme, de sorte qu'elle n'intervient d'aucune façon dans l'étude du corps, qui obéit aux critères de la mécanique.

Certains disciples de Descartes en tireront la conséquence. Ils se sont libérés d'un concept devenu inutile pour l'intelligence du corps et ont abouti au matérialisme.

Dans la mentalité culturelle actuelle, c'est la problématique cartésienne qu'évoque le mot âme.

La connaissance

4. Au regard ouvert sur la réalité des vivants, se présente, avec le règne animal, un nouveau *fait premier*, celui de la *connaissance*. Avec la connaissance, nous nous trouvons en face d'un type de relation entre les êtres, différent des processus, qui peuvent d'ailleurs être présupposés, d'action et passion, d'action et réaction.

La connaissance est une forme de présence – la présence de l'objet connu dans le sujet connaissant. La chose qui est objet de connaissance subsiste dans sa propre existence et, de surcroît, reçoit en tant même que connue un nouveau mode d'existence dans le connaissant, que l'on nomme existence intentionnelle. Précisons que ce qui est connu, ce n'est pas l'image ou le concept de la chose mais, par cette image ou ce concept, la chose elle-même.

L'analyse porte d'abord sur la connaissance sensible, elle distingue les cinq sens externes et quatre sens internes qui sont le sens commun, l'imagination, l'estimative, à laquelle est due comme une ébauche de jugement

portant sur les objets singuliers, et la mémoire. Chez l'homme, à cause de son contact étroit avec la raison, l'estimative est appelée cogitative.

Aux facultés cognitives correspondent les puissances appétitives ou affectives.

La connaissance sensible requiert l'activité d'organes corporels (ajoutons aujourd'hui celle, déterminante, du cerveau).

Le concept

5. Nous rencontrons ici un autre *fait premier*, celui qui constitue l'animal raisonnable dans sa spécificité: l'intellectualité.

L'identification et l'interprétation de ce fait a donné lieu à de nombreux débats philosophiques.

Notre accès aux choses singulières présuppose une multiplicité de perceptions. Par le jeu des comparaisons, qui nous permettent de repérer les différences et les similitudes, nous formons des images générales, qui sont issues des similitudes que nous avons perçues entre des êtres singuliers. Les mots du langage expriment ces généralisations, qui appartiennent encore à la connaissance sensible.

Mais n'expriment-ils que cela? Si la réponse était affirmative, la pensée serait la forme supérieure et la plus complexe de la connaissance sensible.

En réalité, l'image générale, en elle-même et au-delà d'elle-même, est porteuse d'un message d'une autre nature: celui de l'intelligibilité.

Car notre connaissance de la chose n'est pas satisfaite par la constatation de son être-là, elle vise à comprendre ce qui la constitue dans son identité, ce qui fait qu'elle est ce qu'elle est. Elle tend à en saisir l'essence. Cette saisie est telle qu'elle s'applique à tout individu, existant ou possible, d'une espèce donnée; elle ne comporte aucune limitation, elle est de soi universelle.

Il revient à une faculté, distincte de l'imagination et la transcendant, de saisir l'intelligibilité. Cette faculté est l'intelligence ou la raison.[1]

L'intelligence comporte une visualisation d'un autre ordre. Par la médiation du concept, en effet, elle tend à atteindre l'essence de la chose, ce qu'est cette chose. L'idée d'homme est différente de l'image générale de l'homme. A partir du concept que j'abstrais de l'image, je peux former une définition.

Dans la définition de l'homme, je dois évidemment inclure le concept de corps, mais en l'ayant dépouillé des traits singuliers qu'il possède dans la réalité. La définition de l'homme, si elle est correcte, s'applique à Socrate, mais Socrate n'entre pas dans la définition de l'homme.

[1] Ici nous pouvons prendre ces deux notions pour équivalentes.

Nous reviendrons sur l'abstraction par laquelle l'intelligence dégage le concept de l'image.

L'exemple de la définition de l'homme est celui d'une définition qui inclut la référence à la matérialité.

Mais certains concepts, comme ceux d'être, d'un, de vrai, de bien, etc., n'impliquent pas de soi la matérialité.

En parlant de l'intelligence, nous avons parlé de la dimension spirituelle qui distingue l'homme, l'animal raisonnable, des autres animaux, et le constitue dans un ordre propre, qui transcende l'ordre de la pure animalité, bien que par son corps il appartienne lui aussi à l'animalité.

L'abstraction

6. Sur la relation entre l'image et le concept les opinions philosophiques sont divisées: elles vont de l'identité à l'opposition. Nous suivons ici saint Thomas dans sa lecture d'Aristote.

L'objet propre de l'intelligence humaine, celui vers lequel elle se porte en premier et directement est la quiddité des êtres de la nature, qui sont des êtres matériels. Par quiddité (du pronom *quid*), l'on distingue ce qu'est la chose, son identité.

La forme qui donne aux êtres de la nature leur configuration intelligible existe dans les choses singulières. Ce sont ces choses singulières qui se présentent aux sens externes. Puis les données des sens externes accueillies par les sens internes qui les soumettent à une première élaboration aboutissent à une image (*phantasma*) de la chose. Celle-ci, grâce à la perception des similitudes et des différences, trouve sa place dans une classe déterminée d'objets. A l'audition du mot homme, par exemple, surgit en moi une image générale de l'homme dans laquelle se retrouvent toutes les images singulières d'individus humains que j'ai perçues ou reçues.

Ainsi la connaissance sensible nous donne la première présence des choses, elle porte sur des existants.

Cependant, l'image générale n'est pas l'idée. C'est par l'abstraction que l'intelligence dégage cette dernière et visualise la chose à son propre niveau qui est celui de l'intelligibilité.

L'intelligibilité des choses est contenue *en puissance* dans la connaissance sensible que nous en avons. La tradition aristotélicienne et thomiste, qui distingue ces deux formes de connaissance, sensible et intelligible, ne les oppose pas ni ne voit entre elles une coupure dualiste.

Les qualités sensibles saisies par la sensation sont des qualités des choses qui sont et l'être est l'objet de l'intelligence. Celle-ci est ouverte à l'être dans son amplitude universelle, mais il est saisi par nous d'abord dans les quiddités physiques.

L'être est à la fois un et intrinsèquement différencié, il est analogique. En d'autres termes, il est dans chaque être ce qui le constitue en propre et le distingue des autres êtres.

L'analyse de la connaissance intellectuelle conduit à distinguer, à l'origine du processus intellectif une double instance. Sur la nature et le mode de fonctionner de ces deux instances, le Moyen-âge a connu de vifs débats. Thomas d'Aquin a ainsi consacré plusieurs écrits importants au sujet. Il y développe une critique approfondie des commentateurs arabes d'Aristote. On touche ici des questions de fond qui mettent en jeu l'identité même de la personne humaine.

La double instance correspond à deux facultés. L'intelligence qui connaît, appelée intellect possible, reçoit la *species* intelligible par laquelle elle connaît et se dit à elle-même la chose connue.

Mais ce processus présuppose l'action antérieure d'une faculté appelée intellect agent. C'est cette faculté qui *abstrait*.

Abstraire signifie extraire et aussi dégager en dépouillant. Ce qui est dégagé par l'intelligence c'est ce qui constitue formellement la chose, mais dépouillée des notes individuelles que la chose possède dans la réalité. Il existe des corps singuliers, j'en dégage l'idée de corps. Ce dépouillement implique un certain appauvrissement, qui est la rançon de notre accès à l'intelligibilité.

Il existe divers degrés d'abstraction. Dans notre connaissance des êtres physiques, nous dégageons une idée comme l'idée de corps en abstrayant des notes individuelles. Ultérieurement, du monde des corps, je peux abstraire un aspect, celui de la dimension avec ses mesures et ses figures ainsi que des nombres, en faisant abstraction de la nature des objets numérés. En troisième lieu, certains concepts comme l'être, la vérité, la bonté, dans leur contenu ne comportent aucune référence à la matière et au monde des corps.

Le premier temps de l'action de l'intellect agent est la *conversio ad phantasmata*, l'acte de se tourner vers les images d'où il abstrait, en l'élevant à son propre niveau, c'est-à-dire en l'actualisant dans sa propre lumière, l'objet donné d'abord par et dans le sens et qui contient son intelligibilité en puissance.

Un double mouvement

7. La *conversio ad phantasmata*, avons-nous dit, est un premier temps, qui implique un appauvrissement. Mais avec elle, le processus n'est pas achevé. Prenons un exemple tout simple, soit les deux propositions: Socrate est un homme; cet homme est Socrate.

Ce que nous avons vu jusqu'ici permet de comprendre aussitôt la première de ces propositions: l'individu que voici possède la nature humaine ou appartient à l'espèce humaine.

La seconde proposition peut se traduire: cet individu qui possède la nature humaine est cette personne qui a nom Socrate, ou, si l'on veut, mon ami Socrate. En disant ami, je souligne le caractère positif de la singularité, l'unicité de la personne, ce qui introduit à la possibilité d'un rapport avec cette personne d'un nouveau type, dans lequel intervient également l'affectivité.

Cela présuppose un mouvement pour ainsi dire symétrique à la *conversio in phantasmata*, un retour, un reflux, à partir de l'idée dégagée par elle-même, vers la singularité et la richesse du concret, qui avait été écartée dans un premier temps, et se retrouve maintenant assumée dans la lumière de l'intelligence.

L'intelligence connaît l'essence des choses qui, dans la réalité, exercent une existence singulière. Aussi bien, pour mener à terme son opération, l'intelligence a besoin du concours des sens, qu'elle élève dans sa propre lumière, trouvant son achèvement et sa perfection dans la saisie du réel existant.

La spiritualité de l'âme humaine

8. L'âme humaine, qui est la forme du corps est subsistante; elle subsiste par soi, son exister ne dépend pas intrinsèquement du corps qu'elle informe. L'analyse philosophique de Thomas est ici, dans le respect des distinctions, en consonance avec la doctrine de la foi.

C'est à partir de ses activités qu'une substance, sujet premier de ces activités, nous est connue. *Agere sequitur esse*, de l'être découle l'agir.

Pour connaître la nature de l'âme humaine, il est donc nécessaire de partir de ses activités et, plus précisément, de celles qui lui sont spécifiques, comme sont la connaissance intellectuelle et le vouloir.

La connaissance sensible comprend à titre intrinsèque l'activité d'un organe corporel, à la différence de ce qui se passe avec la connaissance intellectuelle.

La relation de cette dernière au corps est autre. Les images, dont l'action abstractive extrait les concepts, sont présupposées à la visualisation intellectuelle mais n'entrent pas dans sa texture.

Une activité déterminée et la faculté dont elle émane directement sont homogènes au sujet qui est à leur principe. Ce sujet ne peut pas être d'un niveau ontologique inférieur à celui des actes dont il est la racine. L'activité intellectuelle et son sujet sont irréductibles à l'activité et à l'âme sensitives.

En tant donc que l'âme humaine est principe d'une activité spirituelle, elle n'est pas soumise à la loi de la génération et de la corruption, caractéristique du monde des corps, c'est-à-dire des êtres composés de matière et de forme. C'est parce qu'elle est au principe d'actes qui transcendent ceux du composé corporel qu'elle n'est pas soumise à la corruptibilité. Elle est subsistante.

Ici l'anthropologie de saint Thomas a bénéficié de la stimulation de la doctrine judéo-chrétienne sur les anges, qui sont de purs esprits.

Dans la hiérarchie des êtres l'âme humaine occupe une position paradoxale. Elle peut être considérée soit un titre de sa spiritualité soit en tant qu'elle est forme du corps.

Dans la *Somme contre les Gentils* (II, c. 68) saint Thomas, qui cite le pseudo-Denys et le livre *de causis*, considère ce qu'il appelle «l'admirable connexion des choses», l'échelle des êtres, sans nier les différences spécifiques, révélant une continuité, puisque nous voyons que toujours ce qui est au dernier degré d'un ordre supérieur touche à ce qui occupe le sommet de l'ordre qui lui est inférieur. Ainsi le corps humain touche la plus humble des substances intellectuelles, qui est l'âme humaine, dont le mode humain d'intellection (*modus intelligendi*) nous prouve la situation au fond de la hiérarchie des esprits. Comme le dit le livre *de causis*, l'âme humaine est «comme un horizon et une frontière entre les substances corporelles et les substances incorporelles», en tant qu'elle est une substance incorporelle, qui cependant est forme du corps.

Il occupe la dernière place dans la hiérarchie des esprits, il est au sommet et au couronnement de l'évolution des vivants. C'est ce que signifie sa définition: *animal rationale*. En conséquence l'anthropologie ressortit à la fois à la métaphysique et à la philosophie de la nature et aux sciences expérimentales.

L'âme humaine, dans ce qui fait sa spécificité, c'est-à-dire en vertu de sa spiritualité, n'est pas soumise à la corruption, comme elle n'est pas le fruit du pur processus biologique de génération. Elle est incorruptible et directement créée par Dieu.

L'objet propre de notre intelligence est, nous l'avons vu, la quiddité des choses sensibles, avec lesquelles, de par notre nature, nous sommes de plain-pied et en contact direct. A partir de la connaissance de ces choses nous accédons à la connaissance des réalités qui les dépassent et qui possèdent une plus riche densité ontologique. C'est la raison pour laquelle, pour exprimer cette positivité supérieure, nous usons de formules négatives. Nous parlons ainsi de l'immatérialité de l'âme, de son incorruptibilité, de son immortalité, pour indiquer que la matière n'intervient pas dans sa constitution et qu'elle appartient à un ordre supérieur, qu'elle n'est pas soumise à la loi de la corruption et de la mortalité.

Le Je pense

Lors de la vogue du structuralisme, la formule *ça pense* a connu un certain succès. La pensée était représentée par là comme l'activité d'un système autorégulé et autopropulsé (une "structure") dont notre esprit serait le lieu et le témoin passif.

La formule tenait de la provocation. Elle s'opposait frontalement au *je pense* qui est une affirmation déterminante de plusieurs philosophies par ailleurs éloignées les unes des autres. En effet, la conscience de soi accompagne les actes de la pensée, entendue au sens le plus large. C'est à moi comme sujet que j'attribue spontanément les actes de connaissance ou les mouvements affectifs qui se présentent dans le champ de ma conscience. Ce n'est pas l'œil qui voit, ce n'est pas le cerveau qui ressent ou qui pense. C'est toujours le sujet, qui au moyen de ces organes, perçoit ou élabore images et idées, au sens que nous avons dit.

Ce sont là des données premières qui ont retenu la réflexion philosophique. Celle-ci, dans son analyse des formes de la causalité, a ainsi dégagé, distincte de la cause principale et sous sa mouvance, le concept de cause instrumentale. Dans notre société technicienne, l'instrumentalité est un fait d'expérience quotidienne. Mais le concept n'est pas restreint aux objets techniques. Il est d'une application plus élémentaire et plus large. Son usage est requis en métaphysique et en théologie. L'instrument est une cause qui n'agit que sous la motion de la cause principale, mais dont l'usage est nécessaire à cette dernière pour qu'elle atteigne sa fin, et qui marque l'effet de sa modalité.

Les organes des vivants entrent dans la catégorie de la cause instrumentale.

Si l'on n'est pas attentif à cette donnée, on court le risque, à son insu, de se trouver enfermé dans la logique du: *ça pense.*

Dans la brève esquisse qui précède, je n'ai pas parlé de ce point, pourtant essentiel, qu'est le langage. Celui-ci nous offre un système de signes qui nous permettent de communiquer entre nous idées et sentiments. Mais le langage nous est également nécessaire pour penser en nous-mêmes. Les mots et les constructions grammaticales de soi appartiennent au domaine des *phantasmata*, objet de la *conversio* et du *reditus;* ils renvoient à un signifié d'un ordre supérieur, celui de l'esprit (la *mens*). Par les mots et les phrases nous pouvons échanger des conceptions qui ressortissent à l'ordre conceptuel. Il faudrait, à ce propos, discuter certaines thèses des récentes philosophies du langage qui font du langage l'ultime instance. En réalité, dans le langage est attestée l'unité du sujet humain et la nature de l'âme spirituelle qui est en même temps forme du corps. En tant qu'elle est spirituelle, elle est directement créée par Dieu.

Developmental Sources of Social Divisions[1]

Elizabeth S. Spelke

Background

My research has aimed for more than 30 years to shed light on human knowledge through study of its origins and development. That study has focused on the fundamental capacities of human infants and young children to reason and learn about the material world of objects, the abstract concepts at the foundations of science and mathematics, and the social world of people. The primary goal of this research, like that of the present workshop, is to shed light on human nature. I believe that we can learn a great deal about our mature selves by contemplating young developing minds, especially in the context of broader studies of human evolution (through systematic comparisons across species) and variability (through systematic comparisons of people living in different cultures and circumstances). To introduce my turn to the question of prejudice, I begin by outlining what I believe we have learned from infants about human nature.

First, infants understand some things but not others. For example, young infants represent objects and expect them to persist and move on contact, but they have no consistent expectations about the behavior of shadows or sand piles. Infants' understanding has been revealed most clearly in five domains: They make coherent, interconnected sets of inferences about inanimate objects and their motions, animate agents and their goal-directed actions, numbers and their relations of ordering and arithmetic, places in the navigable terrain, and geometrical forms and their relations of congruence and scaling (Spelke & Kinzler, 2007; Spelke, 2011). Of course, infants' grasp of each of these entities is highly limited, compared to that of older children and adults. For example, infants represent numbers approximately but not exactly (e.g., Izard, Sann, Spelke & Streri, 2009), and they determine their own spatial position by recording the distances and directions of surrounding surfaces but not the lengths of those surfaces or the angles they form (Lee, Sovrano & Spelke, 2012). Interestingly, young infants' knowledge

[1] I thank Kristin Shutts, Katherine Kinzler, Kristina Olson, Talee Ziv, Gaye Soley, Larisa Heiphetz and Kara Weisman for discussion of the issues addressed in this chapter, and Conisha Cooper for help with preparation of the manuscript. Supported by a grant from the National Institutes of Health (HD-23103).

appears to be shared by a host of other animals from primates to birds and even fish, who show the same abilities and signature limits. For example, chicks represent objects and places much as human infants do (e.g., Chiandetti & Vallortigara, 2011; Lee, Spelke & Vallortigara, 2012).

Infants' cognitive competences are not scaffolding to be thrown away as new capacities emerge. Instead, they are enduring systems of knowledge that continue to function in children and adults, who show the same abilities and limits as infants when they are tested under conditions that prevent the use of later-developing, symbolic skills. When adults must compare sets of objects on the basis of number or perform mental arithmetic under conditions that preclude counting or symbolic calculation, we show similar abilities and limits to those of infants (e.g., Barth, Kanwisher & Spelke, 2003). Moreover, the systems of knowledge that emerge in infancy serve as foundations for later developing uniquely human cognitive skills. For example, the core system of number supports children's learning and adults' performance of symbolic mathematics (e.g., Gilmore, McCarthy & Spelke, 2010; Halberda, *et al.*, 2012), and the core system of place representation supports children's learning to use maps and other spatial symbols (Dillon, Huang & Spelke, in review).

Nevertheless, adults and children all over the world go beyond the limits of these core systems by means of a universal, uniquely human process. Using symbol systems, especially language, we combine the representations delivered by our early developing cognitive systems so as to form new systems of knowledge. For example, infants at the end of the first year begin to combine their representations of objects, actions, and visual forms to create a uniquely human, productive system of knowledge of object kinds: a system that underlies the explosive development of tool use and artifact concepts in the second year of life (Xu, 2009). Moreover, 4-year-old children combine kind representations with their core representations of number to form the uniquely human system of knowledge of natural number, with its associated skills of counting and symbolic arithmetic (Carey, 2009; Spelke, 2000). Recent evidence suggests that older children combine their core geometric knowledge of places and of forms to construct abstract Euclidean geometry (Spelke, 2011; Dillon *et al.*, in review). More speculatively, children may combine their knowledge of living agents and their actions with a sixth system of knowledge, focused on the social world, so as to create uniquely human systems of cooperation and moral evaluation (Spelke, 2010).

These findings shed light on our common humanity. People all over the world, regardless of our specific beliefs and experiences, create our cultures and societies, beliefs and values, with the same tools, upon the same foundations. We have essentially the same interest in and orientation toward the

material world, the living world, and the abstract world of mathematics. Evidence is beginning to suggest, moreover, that people share a common view of the social world. When we go beyond intuition, moreover, we use the same productive cognitive capacities to extend our understanding and create systems of knowledge from formal mathematics to morality. A rich, shared nature unites us.

From this conclusion comes the puzzlement that led me to consider the developmental roots of human social divisions. All over the world, human adults appear to focus more on our differences than on our common humanity. History and contemporary life are marked by potent divisions and conflicts between human groups. The bases for these divisions are diverse, including race, ethnicity, religion, language, national identity, and social status. In all these cases, people go beyond their families and immediate communities and identify and associate with larger groups of individuals who are not personally known to one another but who share a common race, religion, nationality, or some other attribute. In some cases, identification and association with a large group is accompanied by antagonism toward other groups of individuals who differ on one or more of these dimensions. I do not find it surprising that humans around the world make and honor personal commitments to known others, helping family and friends and defending them from harm. Why, however, do we divide the human world into larger social groups, bringing to these groups similar patterns of commitment and conflict?

This question is especially pressing at the present time, because our species faces momentous problems that can only be solved by setting aside our differences and acting collectively to address problems that threaten our survival. We must stop the degradation of our planet, confront global threats to health, and defuse political and economic conflicts that could have worldwide repercussions in this nuclear age. To accomplish any of these tasks, I believe we need a better understanding both of the psychological forces that predispose us to create large-scale social divisions and of the cognitive resources that we can harness to overcome or manage them. Just as studies of infants and children have shed light on the nature, sources, limits, and resilience of our conceptions of objects, living beings, and mathematics, I hope that studies of infants and children will shed light on the nature, sources, limits and strengths of our conceptions of the social world.

Social preferences in infancy

It is evident to the most casual observer that infants, from birth, are bathed in social experience. Given a choice, infants would almost always

rather engage with people than with any other objects or events. Within the first few days of life, infants recognize the faces of their caregivers and the sounds of their speech (see Mehler & Dupoux, 1994). They detect when another person is looking at them and respond with heightened attention (Farroni, Simion, Csibra & Johnson, 2002). If a person gazes at them and then makes a social overture (a facial expression, a vocal exclamation, or gesture of the hand), infants tend to reproduce that gesture (Meltzoff & Moore, 1977), mirroring the expressions of their social partners much as do adults (Chartrand & Bargh, 1999).

In the first months of life, infants also begin to distinguish among people whom they encounter for the first time, based on their appearance. Research from a number of laboratories has probed 3-5 month old infants' reactions to four human social distinctions that have received high attention from social psychologists, and that are visually marked on the human face: distinctions of attractiveness, age, gender and race. In these studies, infants are presented with pairs of photographed faces of unknown people that vary on one of these dimensions, side by side, and their looking times to each face are measured and compared. When two faces differ in age or attractiveness, infants tend to look longer at the younger or more attractive one (Brooks & Lewis, 1976; Langlois, Ritter, Rogman & Vaughn, 1991). When two faces differ in gender, infants looked longer at face of the same gender as their primary caregiver (Quinn *et al.*, 2002; Ramsey, Langlois & Marti, 2005). When two faces differ in race, infants look longer at faces whose race matches that of their families and community (e.g., Bar Haim, Ziv, Lamy & Hodes, 2006).

There are reasons to doubt, however, that these early predispositions are roots of later social divisions and conflicts. First, looking preferences need not be a sign of social preferences. Instead, they may reflect effects of experience on perceptual skill: infants may look longer at faces that are more familiar because those faces are easier to process. Second, the social divisions that fuel wars and other conflicts almost always cross-cut distinctions of attractiveness, age, and gender, and they usually divide human groups more finely than does race, along lines that are difficult or impossible to discern simply by looking at a face. Do infants distinguish between unfamiliar people on the basis of any information that might connect to our mature propensities to divide the social world into internally cooperative and externally competitive groups?

Our first studies, conducted by Katherine Kinzler, focused on a dimension to which we knew infants were sensitive from birth: language and accent. In one study, Kinzler presented 5-month-old infants with

videotaped events showing two unfamiliar people, looking and speaking to the camera as if they were speaking to the infant, in alternation. The two speakers in fact were bilingual in English and Spanish, and each spoke in one of those languages to the infants, who lived in the U.S. in monolingual English-speaking families. Building on the visual preference methods just described, Kinzler tested infants' looking preferences between these two speakers both before and after they addressed the infant: At the beginning and ending of each experiment, the two people stood silent and smiling, side by side, and infants' looking times to each of them were compared. Infants looked equally at the two people at the outset (before they spoke) and during the speaking episodes (watching each person throughout the time that she addressed the infant). At the end of the study, however, infants looked longer at the person who had previously spoken in their native language (Kinzler, Dupoux & Spelke, 2007). Infants' preference for the native language, observed soon after birth (Mehler *et al.*, 1988), here led to a preference for a silent, socially engaging person who previously spoke in that language.

In subsequent studies, infants showed this preference when their native language was paired with faces but not with other visible objects. Moreover, the preference was observed with people whose speaking movements were accompanied by native-accented speech but not by people whose speaking movements were accompanied by familiar or novel inanimate sounds. The preference also was not observed when the videotaped faces and voices were presented in reverse, producing auditory stimulation with the spectral and gross temporal properties of speech, but that does not sound like speech to adults and is not processed as speech by infants (Dehaene-Lambertz, Dehaene & Hertz-Pannier, 2002). Above all, the preference was observed not only when two people spoke to infants in different languages but when both spoke in the infant's native language, with different accents. In this experiment, separate groups of infants in the U.S. and in France were presented with videotaped events depicting the speech of a native speaker of American English and a native speaker of French. Each speaker addressed the monolingual American infants in English and the monolingual French infants in French; infants therefore heard only their native language, spoken either with a native or foreign accent. The infants in both countries showed high and equal looking at the two speakers during both the initial silent presentation and the speaking episodes. After the speaking ended, however, they looked longer at the silently smiling person who had previously addressed them in their native accent. Language and accent consistently influenced infants' looking preferences.

Research by Gaye Soley revealed a further distinction that modulates young infants' looking preferences between unfamiliar people: the songs that they sing (Soley, 2012). Using Kinzler's method, Soley presented 5-month-old infants with videotaped events in which two speakers of their native language sang different songs to the infant. One person sang a song that the infants' parents reported was familiar to them, whereas the other person sang a song of matched rhythm that was unfamiliar (in different studies, the unfamiliar song either had a simple tonal melody or a more complex, atonal melody). Infants looked equally at the two people both during the initial preference test and during the singing, which elicited high attention regardless of the song. In the final test, however, infants looked longer at the person who had previously sung the song that was familiar to them. Interestingly, no such preference was found in a third study, in which the two people sang unfamiliar songs whose melodies were either tonal or atonal. Infants showed looking preferences between two now-silent people only when one person had sung a song that they knew.

Thus, infants show looking preferences for faces on the basis of race, gender, language, accent, and song. In all of these cases, infants look longer at the face whose properties are more familiar. But are any of these looking preferences indicative of social preferences? Moreover, are these reactions found only in infants, or do they endure over development? To address these questions, I turn to studies of older children and to more direct measures of social preference and social engagement.

Preferences of young children for speakers of their language

Kinzler and her collaborators have studied the effect of language on social preferences using diverse methods at ages ranging from 10 months to 6 years. I begin with a study of 10-month-old infants. At the end of the first year, infants begin to share attention to objects and with their social partners (Tomasello, 2008), and object offerings begin to have social meaning. Whereas younger infants respond to toy offerings based only on the properties of the toys, 10-month-old infants assess as well the properties of the person who is offering a toy to them, as if the toy offering were a social overture. Building on these findings, Kinzler and Emmanuel Dupoux, together with Justin Halberda, developed a new method to assess infants' social preferences, focusing on their selective acceptance of toys offered by two different people (Kinzler *et al.*, 2007).

Separate groups of French and American 10-month-old infants were presented with the same videotaped events in which native speakers of French and English alternately spoke to an infant in the speaker's native

language. Then the two people appeared, side-by-side and silently smiling, and each held up an identical toy and offered it to the infant. As each person extended her toy toward the infant, real versions of the two toys moved into view in front of the video images and came to rest in front of the infant, who was allowed to reach for them. French infants reached primarily for the toy offered by the person who had previously spoken in French, and American infants reached primarily for the toy offered by the person who had previously spoken in English. Thus, infants preferentially engaged with the person who had addressed them in their native language.

Kinzler and Kristin Shutts next asked whether infants would selectively learn about objects from native speakers. In two studies, infants viewed videotaped events in which speakers of English and French addressed the infant and then endorsed two different toys or foods. Then infants were allowed to choose between the pair of toys or foods for themselves. At 10 months, infants chose a toy of the type that had been recommended by the native speaker (Kinzler, Dupoux & Spelke, 2012). At 12 months, infants chose to eat food of the type that had been recommended by the native speaker (Shutts, Kinzler, McKee & Spelke, 2009). By the end of the first year, therefore, infants selectively favor the endorsements of others who share their language.

For 2-year-old children, Kinzler devised a different method, focused on children's preferential giving of objects to others (Kinzler *et al.*, 2012). Children living in France and the U.S. were taught a giving game, in which they saw two cartoon characters, side by side on a large video screen, with a real box in front of each character. Children were handed a toy and were encouraged to give it to the character of their choice by putting it inside that character's box; when they did so, the character moved happily in response. After giving toys to each character, the characters were replaced by films of the two women speaking in turn to the child in French and English. Then the women appeared together in silence, and children were handed a toy to give to one of them. Like the cartoon characters, each woman responded silently with a happy gesture and smile when given a toy. The children gave toys primarily to the person who had previously spoken to them in their native language. Although both people were silent and smiling throughout the time that the toddlers handled an object, the language they had previously spoken modulated the toddlers' acts of giving.

With Kathleen Corriveau and Paul Harris, Kinzler tested older children's propensity to learn from others who share their accent with a more explicit measure of selective learning (Kinzler, Corriveau and Harris, 2011). Four- and 5-year-old children were presented with two people who spoke to them in their native language, one with a native and one with a foreign ac-

cent. Then a new object was presented, and children were given the choice to ask one of the people about its function. Children preferentially asked the person who had spoken with the native accent. Finally, children viewed two silent films in which each person demonstrated a different function for the object, and they were asked what they thought its true function was. Children preferentially chose the function that had been demonstrated by the native speaker. At the end of the preschool years, therefore, children still learn selectively from those who share their native accent.

Finally, Kinzler and Shutts tested children's social preferences at 5 years of age with a more explicit measure of social preferences (Kinzler, Shutts, DeJesus & Spelke, 2009). American children were presented with still photographs of other children of the same age, gender and race. An experimenter pointed to each photographed child in turn, invited the participant to listen to that child's voice, and then played a short recording of a child's speech. After the child participant heard one target child speak in their native English with an American accent, and the other child speak either in French or in French-accented English, the experimenter asked which child he or she would rather have as a friend. Children reliably chose as a friend the child who spoke in their native language with a native accent. Indeed, they showed as strong a preference for the native speaker when the contrasting speech was French-accented English as when it was French. In both cases, however, it was possible that children's choices reflected a decision to engage with the person whom they could better understand. A third experiment investigated this possibility by presenting American children with two target children who spoke with French accents, one in French and the other in English. When asked whom they understood, the children chose the French-accented speaker of English. When asked whom they preferred to have as a friend, however, the children chose between the two target children at random. Children's selective association with speakers of their native language and accent evidently does not stem from a strategic choice to engage with people whom they can understand. Instead, children show a social preference for native language speakers.

These findings suggest that language is more than a medium of communication and a critical tool for thought. Language carries social meaning for infants and children. That suggestion is reinforced by research by Kuhl, Tsao & Liu (2003), showing that infants learn language primarily in a social context, from a person with whom they are actively and directly engaged. I will return to the social meaning of language later in this discussion, after considering children's developing social preferences between unfamiliar people who differ on other dimensions.

Preferences of young children for members of their race

While completing their studies of language-based social preferences, Kinzler and Shutts used the same methods to investigate children's social preferences between people who differ in race. Kinzler's first studies focused on 10-month-old infants and used the toy choice method described above, this time presenting videotaped events depicting two women who differed in race (White vs. Black). In one version of the study, both women were presented silently throughout the study, smiling at and gesturing to the infant in alternation and then simultaneously offering toys. In a second version of the study, both women spoke to infants in the same, native language. The study was conducted with infants living in families whose members were White, in two communities in which White people predominated. The findings of both studies contrasted with the findings from the studies presenting speakers of different languages. Infants accepted toys equally from the two women, showing no social preference for the person of their own race (Kinzler & Spelke, 2011).

Kinzler next tested for race preferences in two-year-old children from the same communities. Toddlers were taught the giving game with cartoon characters, as in the study of language preferences, and then were presented with the same two women of different races. Unlike toddlers presented with women who spoke different languages, these toddlers gave toys equally to the women of the two races (Kinzler & Spelke, 2011). At two years of age, children still showed no social preferences between two unfamiliar people on the basis of their race.

In a series of studies, Kristin Shutts tested for race preferences in 3- and 4-year-old U.S. children living in White families, using simple, explicit tasks. In one study, children were shown photographs of two target children of different races but the same age and gender (or, in a different condition, two children of different genders but the same age and race), and they were asked whom they would prefer to play with or to invite to their homes (Shutts, Roben & Spelke, in press). In another study, each of two target children differing in race or gender endorsed a different toy, food or game, and participant children were asked which toy, food or game they would prefer for themselves (Shutts, Banaji & Spelke, 2010). By both measures, 3- and 4-year-old children showed preferences for other children of their own gender. This finding is consistent with longstanding findings that children of this age tend to associate with others of their own gender (Maccoby, & Jacklin, 1987), and it shows that children understood and were engaged by these questions. In contrast, children showed no race preferences at 3 years of age. A year later, race preferences emerged on some measures but not

others. Compared to language, race begins to influence social preferences considerably later in development.

Finally, Kinzler and Shutts tested for race preferences in 5-year-old children, using child photographs and the explicit preference method used in their studies of language-based social preferences. In contrast to the younger children, these White children reliably tended to choose target children of their own race as friends, when the target faces appeared with no language. These findings and others (Aboud, 1988; Hirschfeld, 1996) reveal that sensitivity to race develops by the end of the preschool years. Accordingly, Kinzler and Shutts conducted a final experiment that pitted race against accent. A new group of children saw photographs of the same Black and White children, now accompanied by voice clips in which the Black child spoke English with an American accent and the White child spoke English with a French accent. In this study, the White, American participants showed a reliable preference for the Black, native-accented child over the White, foreign-accented child. Accent trumped race in guiding the social preferences of these children.

Because all of the above studies were conducted on children living in primarily White, monolingual communities, further experiments by Shutts and Kinzler investigated the generality of their findings by testing multilingual children in South Africa, a country with 11 official languages, whose population was segregated, until recently, into four distinct racial groups. Contemporary South African children live in rich multilingual environments; the experience of encountering speakers of other languages therefore is far more familiar to them than to most American children. Because of the country's history of racial apartheid, moreover, children might be expected to show heightened awareness of race. Nevertheless, children in South Africa showed social preferences that were similar to those of their American counterparts. Like children in the U.S., South African children showed reliable preferences for speakers of their native language, relative to speakers of French, a language that is not native to South Africa (Kinzler, Shutts & Spelke, in press). Moreover, South African children of three different racial groups showed little evidence of favoring members of their own racial group (Shutts *et al.*, 2011). Interestingly, both language preferences and race preferences showed some effects of social class: South African children tended to prefer other children whose language or race suggested greater wealth or higher status (see also Olson, Shutts, Kinzler & Weisman, 2012). In South Africa as in the U.S., however, shared language was more powerful than shared race as an influence on children's social choices.

These findings offer a different perspective both on studies of young in-

fants' looking preferences between different faces and on the theory that social preferences are rooted in a predisposition to prefer that which is familiar. Young infants, we saw, look longer at faces of the more familiar race as well as at people who previously spoke in a more familiar language and accent. These different dimensions of familiarity do not, however, appear to have the same social meaning. Although White faces are more familiar than Black faces to the White infants in these studies, infants and young children show no social preference for people of the more familiar race: they do not accept toys more readily from same-race people, they do not place greater trust in the endorsements that same-race people give to objects or foods, they do not offer more gifts to same-race people, and, at 3 years, they do not express a greater desire to befriend same-race people. In contrast, children prefer speakers of their native language by all these measures. Their social preferences between unfamiliar individuals seem not to stem from a general tendency to orient to the familiar, but from something else.

What propels these children's social preferences? Here I focus on one possible reason for the power of language to convey social distinctions. Languages, dialects, and accents are learned from other people over the course of human social interactions. As I noted, infants do not readily learn languages presented outside a social context (Kuhl, *et al.*, 2003). Until the last century, moreover, people had no opportunity even to hear language from non-social sources such a radio: languages were produced only by living people, and usually only when those people interacted with one another. Thus, a person's language and accent depends on the language and accent of the people with whom he or she has directly engaged, over the course of a lifetime. Language and accent are markers of one's social history. When an infant or child encounters another person who speaks like the members of his or her family and community, she may infer that a social chain connects this person to others who speak in the same way. Indeed, prior to the last century, a child could safely infer that a social chain of some length connected such a person directly to herself and her family.

If children are predisposed to favor unknown people who speak their language because such people are likely to know people that the child knows, then language may not be the only factor that modulates young children's social preferences. Children might also be predisposed to favor unknown people who share beliefs or practices that are learned from other members of their community. For example, children might show social preferences for others who share their knowledge of music. Like language, contemporary children learn songs primarily from other people who sing to them. Until the last century, moreover, children and adults throughout the

world learned music only from direct contact with other people who sang or played it. We have seen that infants show looking preferences for those who sing the songs sung by others in the infant's social world. Does music carry social meaning for young children, and if so, does its meaning stem from the status of music as a product of shared cultural knowledge?

Preferences of young children for those who share their cultural knowledge

Gaye Soley (2012) attempted to shed light on these questions through studies of the effects of music on the social preferences of 4-year-old children. In her first study, she presented children with photographs of two target children, accompanied by two songs. Both songs had melodies that are typical of western music, but one was chosen so as to be familiar to the participant child, whereas the other was not. Each song was described as the favorite song of one of the target children. Then children were asked which child they would rather have as a friend. These children chose the child whose favorite song was known to them.

Why did children do this? In describing a song as a child's favorite, this initial experiment conveyed both that the target child knew the song and that he or she liked it. Soley conducted three more experiments to tease apart these factors. In each of these studies, the experimenter presented children with photographs of two target children and played just one song. On half the trials, the song was familiar to the participating children; on the remaining trials, it was novel. Then the experimenter said that she had played the song to the two children in the photographs, and she described the reactions of each target child to the song. In one study, she reported that one child knew the song, and that the other child did not know it, but knew other songs. In a second study, she said that one child liked the song whereas the other child did not like it, but liked other songs. In the last study, she said that one child knew the song but did not like it, and that the other child had not heard the song before, but liked it. After each description, the experimenter asked the participant which child he or she would prefer to have as a friend.

Results were clear: children chose to be friends with the target children whose musical knowledge aligned with their own. In the first study, they preferred children who knew the songs they knew, and children who were ignorant of the songs they did not know. In the second study, the participating children did not differentiate between the target children who liked and disliked the songs that were familiar to them. Instead, children tended to like other children who liked any songs, familiar or unfamiliar. And in the third study, participant children preferred children who knew but did

not like the familiar songs, and children who liked but did not know the unfamiliar songs. Like shared language and accent, shared musical knowledge consistently influenced children's social preferences.

These findings suggest that children favor others who share their cultural knowledge. In turn, this finding accords with the hypothesis that children use music and knowledge not to divide the world of strangers into different groups, but rather to determine whether people whom they meet for the first time are likely to be members of their own immediate community, and to know the people that they know. Music and language influence children's social choices, I suggest, because a person who shares the accent or music of one's friends and family was likely, through most of human history, to be a part of the infant's social world, connected to the infant by a direct chain of social communication and interaction. In the last studies to be described, I turn to our newest work, asking whether this tendency extends from shared language and musical knowledge to shared beliefs. I focus on a powerful set of beliefs that, like language and music, pass from one person to another through social communication: beliefs at the center of formal religions.

Religion is an intense, multifaceted force in the lives of adults and children. Like language and music, it is both universal across human societies and variable across different human cultural traditions. Until recently, children's understanding of religious ideas, and their social preferences for people with differing religious faiths and practices, had received little study. Interest in the developmental psychology of religion recently has begun to increase, but it supports few firm conclusions at this time. My own studies, with Larisa Heiphetz and Mahzarin Banaji, support some tentative suggestions concerning the development of children's social preferences for others who share their religious beliefs and practices.

One series of studies, conducted with 6- to 8-year-old children, used a method similar to that of Shutts and Kinzler (Heiphetz, Spelke & Banaji, in press). Children were introduced to two target children, presented in photographs. One child was described as Christian and was said to engage in a series of Christian practices that are familiar to U.S. children (e.g., painting eggs on Easter). In one study, the other child was described as Jewish and was said to engage in practice that also were familiar to most of the children (e.g., lighting candles on Hanukah). In the other study, the second child was described as Hindu and was said to engage in practices that were not familiar to these children (e.g., lighting lamps on Diwali). Then children were asked whom they would prefer as a friend. Christian children showed no explicit preferences between the Christian and Jewish children, but they showed a weak preference for a Christian child over a Hindu child. Thus,

social preferences based on religion were beginning to appear at this age, distinctly later than preferences based on language or music.

In the next study, with Paul Harris, we asked whether 6- to 9-year-old children would show stronger preferences for those who shared their religion if they were told about target children's beliefs rather than their practices (Heiphetz, Spelke, Harris & Banaji, under review). Moreover, we investigated whether children responded to shared religious beliefs differently than to shared beliefs of other kinds. At the start of this study, we asked children questions so as to assess their own beliefs about matters of fact (e.g., Which do you think is the longer river: the Nile or the Amazon?), of taste (e.g., Which do you think is the better fruit, strawberries or bananas?), and of faith (e.g., When people pray, do you think that God hears them, or do you think that only other people hear them?). Next, we showed children pictures of pairs of target children, one of whom was described as holding the same belief as the child and the other as holding an opposing belief. Children showed reliable preferences for other children who shared their beliefs in all three domains, including the domain of religion. Religion did not appear to hold any special status for the children, however: children chose as a friend the target child who shared their factual beliefs and opinions as reliably as the target child who shared their religious beliefs.

Our studies provide no evidence, thus far, that shared religious beliefs have special importance for children. Moreover, the earlier emergence of a predisposition to endow expressions of belief with social meaning remains unexplored. Nevertheless, this research joins with the studies of shared musical knowledge to suggest that children value others who share their knowledge of the things that people typically learn from other people. For children, shared knowledge may be an indicator of a shared social history.

From family and community to larger social groups

In summary, research on infants and children suggests a developmental progression in children's social preferences between other people whom they do not know. Beginning in infancy, children prefer others who speak in the language and accent, and sing the songs, of the people they know best. At 4-5 years of age, children begin to prefer others who share the race of the people they know best. And by 6-9 years of age, they prefer others who share their beliefs in a range of domains. What aspects of our social nature might explain these findings?

I suggest that all these findings stem from a predisposition to favor unfamiliar others who are likely to be socially connected to the infant's own family and community. In ancestral environments, only other people who

knew members of a community would have spoken in the dialect and accent of that community. Moreover, only such people were likely to know the songs sung by members of that community, the stories and myths recounted by members of community, or the factual knowledge that members of that community gained through their collective experience. Shared language, music and beliefs may be socially meaningful to children because they have served throughout most of our species' history to indicate a direct social connection between the new, unfamiliar people that we encounter and the known, trusted members of our families and their friends. Shared language and music may gain social meaning for children earlier in development than do shared beliefs or rituals, because children begin to learn their native language and songs early in the first year, before they comes to master the community's beliefs.

As children grow, they may discover further attributes that distinguish members of their own community from others. To the degree that communities and social networks are racially segregated, children may learn to use race as an indicator of social allegiances (Cosmides, Tooby & Kurzban, 2003). When learned distinctions of race conflict with our more deeply rooted distinctions based on language, however, language trumps race not only for children, as we have seen, but for adults (Pietraszewski, personal communication).

In a contemporary context, language, music, and other forms of cultural knowledge no longer serve reliably to distinguish those who are, and are not, members of the child's immediate social world. With the development of long-distance travel, colonization, and telecommunications, languages and cultural products have spread far beyond the bounds of any personal social network. With recordings, books and other media, people from widely different cultures and social groups now learn the same languages, songs and stories. Thus, tendencies that could have evolved to allow humans to identify and favor those in their immediate families and communities now will tend to pick out larger groups of unknown individuals. Large-scale social groupings based on language, religion, or ideology might result in part from these tendencies.

Large-scale social groupings that bring together people who share few or no family and direct community ties also might be solidified when institutions use notions of family and community as metaphors. For example, some of the world's most successful and widespread religions use the language of family to unite their members, describing their adherents as brothers, sisters, or children of God. Nations may use the language of community, focusing on cooperation, obedience or mutual personal commitment, to

foster cohesion among their citizens. Much that is good has emerged from such large-scale human groups. Nevertheless, the extension of our natural, early-developing preference for and commitment to our families, friends, and immediate communities to larger groups of unrelated and unknown individuals also brings problems. I end by considering these problems, and the ways in which research on human nature and its development might help to mitigate them.

Core knowledge and human progress

Humans live in an increasingly interconnected world, in which we face common, pressing problems demanding world wide cooperative action. Such problems can best be addressed if people throughout the world recognize our common needs, values and outlook, beneath the divisions of language, race, and belief that direct the interests of one human group against those of others.

Numerous strands of research in the human sciences suggest that this ecumenical stance does not come easily to our species. I have focused on one such strand. A predisposition favor those who speak like the members of our families and community would, in past times, have oriented children toward other people who likely were socially connected to their friends and families. Today, however, it divides the social world into larger groups that sometimes fuel broader conflicts. Others have pointed to our evolutionary heritage as members of small, internally cooperative and externally competitive coalitions, as a source of social propensities that fostered our success in the past but are ill suited to the challenges we face in our contemporary, interconnected world. Can our human minds, shaped by our common history and prehistory in a very different social setting, rise to the challenges that we face today? Research on children's cognitive development suggests some grounds for optimism.

Although human knowledge builds on cognitive systems that emerge in infancy and were shaped by human evolution, the knowledge that we gain has served, again and again in our intellectual history, to carry us beyond these core systems. By constructing new systems of knowledge, spanning several core systems, we can overcome some of the limits of these systems and correct some of the misconceptions to which they give rise.

Consider, for example, historical changes in human conceptions of physical objects. Until the middle ages, conceptions of the physical world built on intuitions that adults share with human infants, and that spring from two core systems for representing manipulable objects and navigable places. According to the first of these systems, objects move only on contact with

other objects: they do not affect each other's motion at a distance. According to the second of these systems, objects and living beings are supported by the ground: a planar surface over which we move and within which we locate ourselves and other objects. These notions began to be overturned, however, when ancient explorers and astronomers discovered that the earth is not flat but round, and when more modern scientists discovered that the earth is a planet orbiting around the sun in response to its gravitational pull. Both the notion of movements on the earth as planar displacements and of causal interactions precluding action at a distance were supplanted by the development of Newtonian mechanics.

The pull of core conceptions still can be felt in children's learning about the physical world. Studies of elementary school children's conceptions of the shape of the earth provide an example (Vosniadou & Brewer, 1992). When such children are asked to describe the shape of the earth in words, they typically report that it is round, but what does this answer mean? Vosniadou and Brewer asked children to draw the earth or model its shape with clay, and to indicate where people stand upon it. Children inventively produced a variety of objects that one could describe as round, but that had a flat upper surface on which the people stood: a pancake-shaped earth, a flattened or hollow sphere with a horizontal top, or an arrangement of two earths: a round, uninhabited planet in the sky and a flat surface below it on which people stood and moved. Children who had been told that the earth is round assimilated that information to their core conceptions of objects and places, resulting in a variety of ingenious misconceptions. Just as clearly, however, children and adults eventually give up their misconceptions and embrace better ones. Good teachers, aware of both the strengths and the pitfalls of children's early developing intuitive conceptions, foster this process. Later in their education, these students will build on their conceptions of manipulable objects, navigable places, and number to embrace systems of knowledge, such as classical mechanics and its successors, that unify these systems and overcome some of the conspicuous misconceptions that arise spontaneously from core systems of knowledge.

As we are inclined to conceive of the earth as flat, we are inclined to conceive of the people who share the earth with us as importantly different from one another, depending on their history, culture, language, and systems of belief. Research in cognitive science suggests, however, that intuition exaggerates these differences, and that humans throughout the world share the same fundamental conceptions, values, and concerns.

Over the course of human history, we have been slower to recognize our misconceptions of the social world than to recognize our misconcep-

tions of the physical world. Two factors, I suggest, may have slowed the advance of progress in mutual understanding. First, social conceptions, even erroneous ones, tend to act as self-fulfilling prophecies. If the members of two groups each suspect that those in the other group reject and therefore threaten their values, they are apt to act in ways that will confirm the opposing group's suspicions. Second, the human mind is more complex than the apples and planets of classical mechanics. It has taken longer for brain and cognitive scientists to discern its outlines.

At present, however, the study of the human mind has progressed, as the research presented at this workshop attests. Moreover, the fates of all humans are ever more deeply interconnected, increasing the need for, and importance of, this new understanding. I believe that we can use our capacities for cognitive progress to recognize our social misconceptions, and to develop a sense of human nature that is both more accurate and more adequate to our current challenges. I hope that research in the human sciences will be helpful in this regard.

References

Aboud, F. (1988). *Children and prejudice.* Cambridge, MA: Basil Blackwell.

Bar-Haim, Y., Ziv, T., Lamy, D., & Hodes, R. (2006). Nature and nurture in own-race face processing. *Psychological Science, 17,* 159-163.

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition, 86*, 201-221.

Brooks, J., & Lewis, M. (1976). Infants' responses to strangers: midget, adult, and child. *Child Development, 47*, 323-332.

Carey, S. (2009). *The origin of concepts.* Oxford: Blackwell.

Chartrand, T.L. & Bargh, J.A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*, 893-910.

Chiandetti, C. & Vallortigara, G. (2011). Intuitive physical reasoning about occluded objects by inexperienced chicks. *Proc. Roy. Soc. B., 278,* 2621-2627.

Cosmides, L., Tooby, J., & Kurzban, R. (2003). Perceptions of race. *Trends in Cognitive Sciences, 7,* 173-178.

Dehaene-Lambertz, G., Dehaene, S. & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science*, *298*, 2013-2015.

Farroni, T., Csibra, G., Simion, F., & Johnson, M.H. (2002). Eye contact detection in humans from birth. *Proc. Nat. Acad. Sci. (USA), 99*(14), 9602-9605.

Gilmore, C. K., McCarthy, S.E., & Spelke, E.S. (2010). Non-symbolic arithmetic abilities and mathematics achievement in the first year of formal schooling. *Cognition*, *115*(3), 394-406.

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D.Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive internet-based sample. *Proc. Nat. Acad. Sci. (USA), 107*(28), 11116-11120.

Heiphetz, L., Banaji, M.R., & Spelke, E.S. (in press). I can't see it, but I know it matters: Religious group membership guides Christian children's preferences for novel characters. *Journal of Experimental Psychology: General.*

Hirschfeld, L. (1996). *Race in the making.* Cambridge, MA: MIT Press.

Izard, V., Sann, C., Spelke, E.S. & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences, 106*(25), 10382-10385.

Kinzler, K.D., Corriveau K.H. & Harris, P.L. (2011). Children's selective trust in native-accented speakers. *Developmental Science*, *14*, 106-111.

Kinzler, K.D., Dupoux, E., & Spelke, E.S. (2007). The native language of social cognition. *Proc. Nat. Acad. Sci. (USA), 104,* 12577-12580.

Kinzler, K.D., Dupoux, E., & Spelke, E.S. (2012). 'Native' objects and collaborators: Infants' object choices and acts of giving reflect favor for native over foreign speakers. *Journal of Cognition and Development, 13,* 67-81.

Kinzler, K.D., Shutts, K., DeJesus, J., & Spelke, E.S. (2009). Accent trumps race in children's social preferences. *Social Cognition, 27,* 623-634.

Kinzler, K.D., Shutts, K., & Spelke, E.S. (in press). Language-based social preferences among multilingual children in South Africa. *Language learning and development.*

Kinzler, K.D. & Spelke, E.S. (2011). Do infants show social preferences for people differing in race? *Cognition*, *119*, 1-9.

Kuhl, P., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proc. Nat. Acad. Sci. (USA), 100.* 9096-9101.

Langlois, J.H., Ritter, J.M., Roggman, L.A., & Vaughn, L.S. (1991). Facial diversity and infant preferences for attractive faces. *Developmental Psychology, 27*(1), 79-84.

Lee, S.A., Sovrano, V. & Spelke, E.S. (2012). Navigation as a source of geometric knowledge: Young children's use of length, angle, distance and direction in a reorientation task. *Cognition, 123,* 144-161.

Lee, S.A., Spelke, E.S. & Vallortigara, G. (2012). Chicks, like children, spontaneously reorient by three-dimensional environmental geometry, not by image matching. *Biology Letters, 8,* 492-494.

Maccoby, E.E., & Jacklin, C.N. (1987). Gender segregation in childhood. In E.H. Reese (Ed.), *Advances in child development and behavior* (Vol. 20, pp. 239-287). New York: Academic Press.

Mehler, J. & Dupoux, E. (1994). *What infants know.* NY: Wiley.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29,* 143-178.

NY: Basic Books.

Meltzoff, A.N. & Moore, M.K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*, 75-78.

Olson, K., Shutts, K., Kinzler, K.D. & Weisman, K. (2012). Children associate racial groups with wealth: Evidence from South Africa. *Child Development,* in press.

Quinn, P.C., Yahr, J., Kuhn, A., Slater, A.M., & Pascalis O. (2002). Representation of the Gender of Human Faces by Infants: A Preference for Female. *Perception*, *31*, 1109-1121.

Ramsey, J.L., Langlois, J.H., & Marti, N.C. (2005). Infant categorization of faces: ladies first. *Developmental Review, 25*, 212-246.

Shutts, K., Banaji, M.R., & Spelke, E.S. (2010). Social categories guide young children's preferences for novel objects. *Developmental Science, 13,* 599-610.

Shutts, K., Kinzler, K.D., Katz, R.C., Tredoux, C., & Spelke, E.S. (2011). Race preferences in children: Insights from South Africa. *Developmental Science, 14,* 1283-1291.

Shutts, K., Kinzler, K.D., McKee, C.B., & Spelke, E.S. (2009). Social information guides infants' selection of foods. *Journal of Cognition and Development, 10,* 1-17.

Shutts, K., Roben, C.K.P. & Spelke, E.S. (in press). Children's use of social cate-

gories in thinking about people and social relationships. *Journal of Cognition and Development*.

Soley, G. (2012). *Exploring the nature of early social preferences: The case of music.* Unpublished doctoral dissertation, Harvard University.

Spelke, E.S. (2000). Core knowledge. *American Psychologist, 55*, 1233-1243.

Spelke, E.S. (2011). Natural number and natural geometry. In E. Brannon & S. Dehaene (Eds.), *Space, Time and Number in the Brain: Searching for the Foundations of Mathematical Thought* (pp. 287-317). Attention & Performance XXIV, Oxford University Press.

Spelke, E.S., Bernier, E. & Skerry, A. (in press). Core Social Cognition. In M.R. Banaji & S.A. Gelman (Eds.), *Navigating the Social World: What Infants, Children, and Other Species Can Teach Us.* Oxford University Press.

Spelke, E.S., & Kinzler, K.D. (2007). Core knowledge. *Developmental Science, 10*, 89-96.

Tomasello, M. (2008). *The origins of human communication*. Cambridge, MA: MIT Press.

Vosniadou, S. & Brewer, W. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*, 535-585.

Xu, F. (2007). Sortal concepts, object individuation, and language. *Trends in Cognitive Sciences, 11,* 400-406.

Can Neuroscience Improve the Brain and Mind?

How Genes and Experience Shape the Human Will[1]

Michael I. Posner, Mary K. Rothbart, Pascale Voelker[2] & Yi-Yuan Tang[3]

In this paper we examine mechanisms underlying *will*. By *will* we refer to the means through which our intentions lead to our thoughts and behaviors. In psychology these mechanisms are variously called self regulation in childhood and self or cognitive and emotional control in adults. In our view all these various names refer to specific brain networks that comprise the organ system involved in attention (Posner, 2012).

During early development, attentional orienting operates in conjunction with the actions of caregivers to provide regulation of behavior. This association underlies the frequent observation that infants and young children are largely controlled by their environment including caregivers. Later the executive attention network comes to dominate self regulation allowing internal goals to guide behavior. In this paper we trace the development of the mechanisms through which volition operates. We consider both the common networks that serve as mechanisms of voluntary control for all people and individual differences in the efficiency of these networks. We stress the joint role of genes and experience, including parenting, in the development of the will and in the ability to modify it through training. Finding the mechanisms of will may further our understanding of the extent to which will is constrained as well as free.

Self Regulation and Attention

Imaging the human brain by use of functional magnetic resonance (fMRI) has revealed brain networks related to specific aspects of attention including obtaining and maintaining the alert state, orienting to sensory stimuli and resolving conflict among competing responses (Posner & Petersen, 1990; Petersen & Posner, 2012).

[1] This paper is based on a lecture at the Vatican Nov. 8, 2012 by the first author. This research was supported by NICHD Grant HD 060563 and Office of Naval Research grant N000141110034 to the University of Oregon.

[2] University of Oregon, Eugene, OR, USA.

[3] Texas Tech University, Lubbock, TX, USA.

The alerting network is modulated by the brain's norepinephrine system and involves major nodes in frontal and parietal cortex. The alert state is critical to high-level performance. Phasic changes in alertness can be produced by the presentation of a warning of an impending target. This leads to a rapid change from a resting state to one of increased receptivity to the target.

The orienting network interacts with sensory systems to amplify information relevant to task performance. The orienting network exerts much of the regulatory control present during infancy and early childhood (Posner, Rothbart, Sheese & Voelker, 2012; Rothbart, Sheese, Rueda & Posner, 2011).

The executive network is involved in resolving competing actions in tasks where there is conflict. This is done both by enhancing activity in networks related to our goals and inhibiting activity in conflicting networks. These controls are effected by long connections between the nodes of the executive network and cognitive and emotional areas of the frontal and posterior brain. In this way the executive network is important for voluntary control and self regulation (Bush, Luu & Posner, 2000; Sheth *et al.,* 2012). Effortful control is a higher order temperament factor assessing self regulation that is obtained from parent report questionnaires (Rothbart, 2011). In childhood, performance on conflict related cognitive tasks is positively related to measures of children's effortful control (Rothbart, 2011). During childhood and in adulthood effortful control is correlated with school performance and with indices of life success, including health, income and successful human relationships (Checa & Rueda, 2011; Moffitt *et al.,* 2011).

There are individual differences in the efficiency of each of the three networks. The Attention Network Test (ANT) was devised as a means of measuring these differences (Fan *et al.,* 2002). The task requires the person to press one key if a central arrow points to the left and another if it points to the right. Conflict is introduced by having surrounding flanker arrows point in either the same (congruent) or the opposite (incongruent) direction. Cues presented prior to the target provide information on where or when the target will occur. Three scores are computed that are related to the performance of each individual in alerting, orienting and executive control. In our work we have used the Attention Network Test (ANT) to examine the efficiency of brain networks underlying attention (Fan *et al.,* 2002). A children's version of this test is very similar to the adult test, but replaces the arrows with fish (Rueda *et al.,* 2004).

Studies have shown moderate reliability of conflict scores, but much lower reliability for the orienting and alerting scores (MacLeod *et al.,* 2010) and recent revisions of the ANT provide better measures of orienting and alerting that may improve these results (Fan *et al.,* 2010). The attentional

networks involve different cortical brain areas (Fan *et al.,* 2005), and scores on the ANT are related to distinct white matter pathways (Niogi & McCandliss, 2009) and structural differences in cortical thickness (grey matter) (Westlye, Grydland, Walhove, Fjell, 2011). The ANT and its various revisions show significant interaction between networks (Callejas, Lupianex & Tudela, 2004; Fan *et al.,* 2010). It is clear that the networks communicate and work together in many situations, even though their anatomy is distinct.

The dorsal part of the anterior cingulate cortex (ACC) is involved in the regulation of conflict in cognitive tasks, while the more ventral part of the cingulate is involved in regulation of emotion (Bush, Luu & Posner, 2000). One way to examine this issue is to image the structural connections of different parts of the cingulate using diffusion tensor imaging (DTI). This form of imaging traces the diffusion of water molecules in along myelinated fibers, and provides a means of examining the physical connections present in the brain. DTI studies have shown that the dorsal part of the ACC is connected primarily to parietal and frontal lobes, while the ventral part of the ACC has strong connections to subcortical limbic areas (Posner, Sheese, Odulas and Tang 2006).

The executive attention network also includes the underlying basal ganglia and adjacent areas of the prefrontal cortex. There is evidence that the anterior insula is involved particularly when switching between tasks (Supekar & Menon, 2012), while adjacent midprefrontal cortex is important during complex decision making (Behrens, 2012). Comparative anatomical studies point to important differences in the evolution of cingulate connectivity between non-human primates and humans. Anatomical studies show the great expansion of white matter, which has increased more in recent evolution than has the neocortex itself (Zilles, 2005). One type of projection cell called the Von Economo neuron is found only in the anterior cingulate and a related area of the anterior insula, two brain areas that are active together even when the person is resting and not performing a task (Allman, Watson, Tetreault, and Hakeem, 2005; Dosenbach *et al.,* 2007). It is thought that von Economo neurons are important in communication between the cingulate and other brain areas. This neuron is not present at all in macaques and its frequency increases between the great apes and adult humans. Moreover, there is some evidence that the frequency of the neuron increases in development between infancy and later childhood (Allman, *et al.,* 2005).

Development of Self Regulation

Resting State Studies

It has recently become common to study the brains of children and adults while they are resting (resting state rsMRI; Raichle, 2009). One of

the brain networks active during rest is the executive attention network involved in resolving conflict and related to parent reports of effortful control (Dosenbach *et al.,* 2007; Fair *et al.,* 2009). Resting state methods can be applied at any age because they do not require a task. Studies have examined how brain networks change with age (Fair *et al.,* 2009; Gao, *et al.,* 2009). Resting state studies have found that during infancy and early childhood most brain networks involve short connections between adjacent areas, but the long connections important for self regulation develop slowly over childhood (Fair *et al.,* 2009; Gao *et al.,* 2009). Development of this system may relate to the achievements in self regulation that we have documented between infancy and 7-8 years of age (Rueda *et al.,* 2004). A study using fMRI for 725 children from 4 to 21 years showed a relationship between the ability to resolve conflict in a flanker task and the size of the right dorsal anterior cingulate in the early years of childhood as well as the connectivity of the cingulate in later years (Fjell *et al.,* 2012).

Longitudinal Study

In our work we found that 7-month-old infants activated the anterior cingulate when they detected an error (Berger, Tzur & Posner, 2006; Wynne, 1994), showing that they have rudimentary executive attention in place, even though parents are not yet able to report on effortful control and infants do not carry out instructed behaviors. It was not until age three that children began to show regulation of their behavior by slowing their next response following an error as adults do (Jones, Rothbart & Posner, 2003).

We conducted a longitudinal study on the development of self regulation during infancy and childhood. The testing began when the infants were 7 months old. Because infants are not able to carry out voluntary attention tasks, we used a visual task in which a series of attractive stimuli are put on the screen in a repetitive sequence (Clohessy, Posner & Rothbart, 2001; Haith, Hazan & Goodman, 1988). Infants orient to them by moving their eyes (and head) to the location. On some trials infants showed they anticipated what was coming by orienting prior to the stimulus. We found (Sheese *et al.,* 2008) that infants who made the most anticipatory eye movements also exhibited a pattern of cautious reaching toward novel objects that predicts effortful control in older children (Rothbart, 2011). In addition, infants with more anticipatory looks showed more spontaneous attempts at self regulation when presented with somewhat frightening objects.

We retested and genotyped the children at age 18 to 20 months and tested them again at about age 4 when they were able perform the ANT as a measure of executive attention. We found that the early regulatory ef-

fects in infancy and at age 2 were correlated with their later orienting network scores rather than their executive network performance in the ANT. In addition, we found that in infancy orienting of attention was related to lower negative and higher positive affect. By age two, orienting was no longer related to affect, but later in childhood and for adults, effortful control is related to lower negative affect (Rothbart, 2011).

These findings led us to the view that the orienting network provides the primary regulatory function during infancy. The orienting network continues to serve as a control system, but starting in childhood the executive attention appears to dominate in regulating emotions and thoughts (Isaacowitz, 2012; Posner *et al.,* 2012; Rothbart *et al.,* 2011). This parallel use of the two networks fits with the findings of Dosenbach *et al.* (2007) that in adults the frontal-parietal network controls task behavior at short time intervals whereas the cingulo-opercular network exercises strategic control over long intervals.

Our general view is that cognitive and emotional control systems arise as part of attentional networks. Our view contrasts with some general views of cognitive development that see self control as arising out of the child's ability to employ language to implement control (Luria, 1973; Vygosky, 1934). While we do not deny that language plays an important role in development, we have observed young children when they attempt to exercise control of dominant responses in Simple Simon type games (Jones *et al.,* 2005). In general, self regulation involved physical actions such as children sitting on their hands or holding one hand with the other, rather than self instruction by language. In addition, the importance of the growth of the right rather than the left ACC in self regulation (Fjell *et al.,* 2012) and association of self regulation with attentional orienting support a separation of control from language. It seems to us that previous work has failed to recognize the separate evolutionary development of attention systems as a basis for control, placing too much emphasis on language as a uniquely human control system.

Physical changes in connectivity

The changes in connectivity during development reported in resting state MRI studies involve functional connectivity based upon correlations between BOLD activity in separated brain areas. Is there evidence of the actual physical changes in the white matter thought to underlie these correlations? Our recent work with adults using DTI has uncovered white matter changes that have similarities to those found in development. Training adults might thus allow us to uncover how the connections de-

veloping during childhood support the changes in self control between infancy and adults.

During development there is a large change in the physical connections between brain areas. The number of axons connecting brain areas increases followed by an increase in the myelin sheath that surrounds the axon and provides insulation. Together these changes result in more efficient connections (Lebel *et al.,* 2012). Fractional anisotropy (FA) is the main index for measuring the integrity of white matter fibers when using DTI.

In our work we studied FA in college students before and after a form of mindfulness meditation called Integrated Body Mind Training (IBMT) in comparison to a control group given the same amount of relaxation training. We found clear improvement in the executive attention network after only five days of training. After two to four weeks of training we found significantly greater change in FA following meditation training than following the relaxation training control in all areas of connectivity of the ACC, but not in other brain areas (Tang *et al.,* 2010).

These alterations in FA could originate from several factors such as changes in myelination, axon density, axonal membrane integrity, axon diameter and intravoxel coherence of fiber orientation and others. Several DTI studies have examined axial diffusivity (AD) and radial diffusivity (RD), the most important indices associated with FA, to understand the mechanisms of FA change (Bennett *et al.,* 2010; Burzynska *et al.,* 2010). Changes in AD are associated with axon morphological changes, with lower AD value indicating higher axonal density. In contrast, RD implicates the character of the myelin. Decreases in RD imply increased myelination, while increases represents demyelination.

In our study (Tang *et al.,* 2012) – we investigated AD and RD where FA indicated that integrity of white matter fibers was enhanced in the IBMT group more than control group. We found that after two weeks there were changes in axonal density but not in myelination. In some areas these changes in axonal density were correlated with improved mood and affect as measured by self report. After 4 weeks of training we found evidence of myelination changes. Since the developmental changes in childhood first involve changes in axonal density and only later myelination, our training may provide changes that are somewhat similar to those found in development. If so, it might be possible to use training to study how physical changes in connectivity alter aspects of control including reaction time, control of affect, stress reduction and other changes found with meditation training.

Genes and Environment

We have pursued two strategies to help understand how genes are related to the individual efficiency of attention networks. One approach involves adults and uses the association of attentional networks with particular neuromodulators (Green *et al.,* 2009; see also Table 1). These associations have led to identification of candidate genes that are expected to relate to each network. The results were summarized by Green *et al.,* (2009) and are shown in Table 1. Since 2009 a number of other results have qualified the view somewhat. It seems clear that serotonin as well as dopamine can influence the executive attention network (Reuter *et al.,* 2007), and that there are interactions between dopaminergic and cholinergic genes at the molecular level that modify the degree of independence between them (Market *et al.,* 2010). Nonetheless the scheme in Table 1 provides a degree of organization and prediction that is often lacking in studies of genetic influence on cognition and behavior.

Table 1. Relating Attention Networks to Dominant Modulators and Genetic Alleles

NETWORK	MODULATOR	GENES
ALERTING	NOREPINEPHRINE	ADRA2A NET
ORIENTING	ACETYCHOLINE	CHRNA4 APOE
EXECUTIVE	DOPAMINE	DRD4, DAT1, COMT MAOA, DBH
	SEROTONIN	TPH2, 5HTT

This table adapted from Green *et al.,* 2008

Some individuals have stronger activations and connectivity than others and are thus better able to exercise the various functions of self regulation. Moreover childhood assessments of effortful control (Moffitt *et al.,* 2011) and self regulation (Casey, *et al.,* 2011) predict performance as adults. How do these differences arise? In part they are due to genetic variations. However, environmental influences and learning can also lead to differences in

efficiency; so experience and genetics are not separate influences but frequently interact. Gene expression, for example, can be altered by the environment in which the genes operate. Genetic differences can also influence the degree to which specific experience is effective in leading to learning (Belsky & Pluess, 2009). Our genes thus influence the degree to which our behavior is altered by experience. This is a far cry from the view of immutability of genes that might underlie the idea that our weaknesses are unchangeable (see also Spector, 2012).

We illustrate the complex interaction between genetic variation and environmental influence with results obtained in our longitudinal study that involve the 7-repeat allele of the DRD4 gene. This allele has been associated with attention-deficit/hyperactivity disorder (ADHD) and the temperamental quality of sensation seeking. Evidence that environment and/or experience can have a stronger influence in individuals with the 7-repeat allele has been reported (Bakermans-Kranenburg and van Ijzendoorn, 2006; van Ijzendoorn & Bakermans-Kranenburg, 2006). Moreover, in one study an intervention that increased parent use of positive discipline reduced externalizing behavior in toddlers with the 7-repeat allele of the DRD4 gene significantly more than for those without this allele (Bakersman-Krannenburg, Ijzendoorn, Pijlman, Mesman, & Juffer, 2008). This finding is important because assignment to the intervention group was random, thus ensuring that the result is not due to something about the parents other than the training.

In our longitudinal study conducted at the University of Oregon, cheek swabs were used to collect DNA samples and genetic variation was identified in twelve genes that had been related to attention in adult studies (Sheese, Voelker, Rothbart, & Posner, 2007). The children had been evaluated when they were 7 months old, and genotyping took place when they returned to the laboratory at about 2 years of age. In addition, parenting quality was examined through observation of caregiver–child interactions in which the children played with toys in the presence of one of their caregivers. Raters reviewed videotapes of the caregiver–child interaction and rated the parents on five dimensions of parenting quality according to a schedule developed by NICHD (1993): support, autonomy, stimulation, lack of hostility, and confidence in the child. According to their scores, parents were divided at the mean into two groups: one showing a higher quality of parenting, and the other a lower quality. Results showed an interaction between parenting quality and variation of the DRD4 gene. For children with the 7-repeat allele, there was a strong influence of parenting quality. Children with the 7-repeat allele and lower quality parenting were high in

impulsivity while those with higher quality parenting were normal in impulsivity. Children without the 7 repeat allele showed normal levels of impulsivity regardless of parent quality. Similar results were obtained for activity level and high-intensity stimulation seeking, which can be combined with impulsivity into one aggregate measure of sensation seeking.

Some evidence suggests that the 7-repeat allele is under positive selective pressure in recent human evolution (Ding, Chi, Grady, Morishima, Kidd, *et al.,* 2002). Why should an allele that has been found to be over represented in Attention Deficit Disorder (ADHD) be undergoing positive selection? We think that positive selection of the 7-repeat allele could well arise from its sensitivity to environmental influences. Parenting provides training for children in the values favored by the culture in which they live. For example, Rothbart and colleagues (Ahadi, Rothbart, & Ye, 1993) found that in Western culture, effortful control appears to regulate negative affect (sadness and anger), while in China (at least in the 1980s) it was found to regulate positive affect (outgoingness and enthusiasm). In recent years, the genetic part of the nature-by-nurture interaction has been given a lot of emphasis. Theories of positive selection in the DRD4 gene have stressed the role of sensation seeking in human evolution (Harpending & Cochran, 2002; Wang, Kodama, Baldi, & Moyzis, 2006). The finding that individual differences in impulsivity may be influenced by the interaction between genetics and parenting style do not contradict this evolutionary emphasis, but suggest a form of explanation that could have even wider significance. If genetic variations are selected according to their sensitivity to cultural influence, this could fit with evidence that the 7 repeat allele is under positive selective pressure and is over represented among the elderly (Grady *et al.,* 2013). It remains to be seen whether the other 300 genes estimated to show positive selection would also increase an individual's sensitivity to variations in rearing environments.

How could variation in genetic alleles lead to enhanced influence of cultural factors such as parenting? The anterior cingulate receives input pertaining to both reward value and pain or punishment, and this information is clearly important in regulating thoughts and feelings. Dopamine is the most important neuromodulator in these reward and punishment pathways. Thus, changes in the response to dopamine could enhance the influence of signals from parents related to reward and punishment. Because the ACC is important in executive attention, we expected that the 7 repeat influence on behavior was mediated by executive attention. However, in the study by Sheese and colleagues (2007), data showed that at two years there was no influence of the 7-repeat allele on executive attention; rather, the gene

and environment interacted to influence the child's behavior as observed by the caregiver. However, the same children at age 4 did show an interaction between the presence of the DRD4 7-repeat allele and parenting quality in determining effortful control and this effect has been replicated in another study (Sheese *et al.,* 2012; Smith *et al.,* 2012). Since effortful control is linked to executive attention, this finding suggests that the executive network could be a mechanism for the widespread effects of Gene x Environment interactions, at least in older children and adults. Thus the DRD4 7 repeat allele may operate through executive attention after age 4 but through some other mechanism before executive attention is sufficiently developed.

There is other evidence that despite changes in the brain and behavior, the DRD4 7 repeat allele may play the same role in adults as in children. This study (Larson, *et al.,* 2010) exposed young adolescents to either high or low levels of alcohol consumption by peers. Adolescents with the 7 repeat allele were more influenced in their drinking by peer behavior than those without the 7 repeat.

We have illustrated the way parenting and other environmental influences may show continuity, despite changes in the underlying control networks, with the story of only one genetic allele. However, a recent selective review of the longitudinal literature (Ronald, 2011) reaches a similar conclusion over a larger number of genes and studies. More data in this area are needed to better understand the mechanisms by which genes operate over the lifespan.

This paper has examined attention and the mechanisms supporting will. We view will as involving regulation of cognitive and emotional systems, and argue that such self regulation is carried out in older children and adults by the executive attention network. We have traced the early reliance of control on the orienting network and seen how during development control shifts to the executive network. We have examined methods of changing the efficiency of self regulation through training as a model for what might happen in development. Finally we examined how genes influence the degree to which training can shape self regulation. We hope that this approach will eventually provide a more complete picture of the origins and neural systems of voluntary control and provide scientific data on the constraints that operate on *will*.

References

Ahadi, S.A., Rothbart, M.K., & Ye, R. (1993). Children's Temperament in the U.S. and China: Similarities and differences. *European Journal of Personality,* 7, 359-378.

Allman, J., Watson, K.K., Tetreault, N.A., & Hakeem, A.Y. (2005). Intuition and autism: A possible role for von Economo neurons. *Trends in Cognitive Science,* 9:367-373.

Bakermans-Kranenburg, M.J., & van Ijzendoorn, M.H. (2006). Gene-environment interaction of the dopamine D4 receptor (DRD4) and observed maternal insensitivity predicting externalizing behavior in preschoolers. *Developmental Psychobiology*, *48*, 406-409.

Bakersmans-Kranenburg, M.J., van Ijzendoorn, M.H., Pijlman, F.T.A., Mesman, J., & Juffer, F. (2008). Experimental evidence for differential susceptibility: Dopamine D4 receptor polymorphism (DRD4VNTR) moderates intervention effects on toddlers externalizing behavior in a randomized controlled trial. *Developmental Psychology, 44*, 293-300, doi:10.1037/0012-1649.44.1.293

Behrens, T.E. (2013). Neural mechanisms underlying human choice in frontal cortex. In Pontifical Academy of Sciences, *Neurosciences and the Human Person*, Proceedings of the Working Group of 8-10 November 2012, *Scripta Varia* 121, Vatican City.

Belsky, J., & Pluess, M. (2009). Beyond diathesis stress: Differential susceptibility to environmental stress. *Psychological Bulletin*, 135:895-908.

Bennett I.J., Madden D.J., Vaidya C.J., Howard D.V., Howard J.H., Jr. (2010). Age-related differences in multiple measures of white matter integrity: A diffusion tensor imaging study of healthy aging. *Hum. Brain Mapp.* 31:378-390.

Berger, A., Tzur, G., & Posner, M.I. (2006). Infant brains detect arithmetic errors. *Proceedings of the National Academy of Science of the United States of America*, *103*, 12649-12653, doi:10.1073/pnas.0605350103

Burzynska AZ, *et al.,* (2010). Age-related differences in white matter microstructure: region-specific patterns of diffusivity. *Neuroimage* 49: 2104-2112.

Bush, G., Luu, P. & Posner, M.I. (2000). Cognitive and emotional influences in the anterior cingulate cortex. *Trends in Cognitive Science*, 4/6:215-222.

Callejas, A. Lupianez, J. & Tudela, P. (2004). The three attentional networks: on their independence and interactions. *Brain and Cognition* 54(3) 225-227.

Casey, B.J. *et al.,* (2011). Behavioral and neural correlates of delay of gratification 40 years later. Proceedings of the National Academy of Sciences USA 108/36, 14998-1503.

Checa, P. & Rueda, M.R. (2011). Behavior and brain measures of executive attention and school competence in late childhood *Developmental Neuropsychology* 36/8, 1018-1032, doi:10.1080/87565641.2011.591857

Clohessy, A.B., Posner, M.I., & Rothbart, M.K. (2001). Development of the functional visual field. *Acta Psychologica*, 106:51-68.

Ding, Y.C., Chi, H.C., Grady, D.L., Morishima, A., Kidd, J.R., Kidd, K.K. *et al.,* (2002). Evidence of positive selection acting at the human dopamine receptor D4 gene locus. *Proceedings of the National Academy of Sciences of the USA, 99*(1), 309-314.

Dosenbach, N.U.F. Fair, D.A. Miezin, F.M. Cohen, A.L. Wenger, K.K. R. Dosenbach, A.T. Fox, M.D. Snyder, A.Z. Vincent, J.L. Raichle, M.E, Schlaggar, B.L. and Petersen, S.E. (2007). Distinct brain networks for adaptive and stable task control in humans, *Proceedings of the National Academy of Sciences of the USA* 104, 1073-1978.

Fair, D.A., Cohen, A.L., Power, J.D., Dosenbach, N.U.F., Church, J.A., Miezin, F.M.,

Schlaggar, B.L., & Petersen, S.E. (2009). Functional brain networks develop from a "local to distributed" organization. *PLoS Computational Biology, 5*, e1000381, doi:10.1371/journal.pcbi.1000381

Fan, J., Gu, X., Guise, K.G., Liu, X., Fossella, J., Wang, H., & Posner, M.I. (2009). Testing the behavior interaction and integration of attentional networks. *Brain and Cog.* 70, 209-220.

Fan, J., McCandliss, B.D., Sommer, T., Raz, M. & Posner, M.I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, 3(14):340-347.

Fjell, A.M. *et al.,* (2012). Multi modal imaging of the self-regulating brain. *Proceedings of the National Academy of Sciences USA published ahead of print, November 12, 2012,* doi:10.1073/pnas.1208243109

Gao, W., Zhu, H., Giovanello, K.S., Smith, J. K., Shen, D., Gilmore, J.H., & Lin, W. (2009). Evidence on the emergence of the brain's default network from 2-week-old to 2-year-old healthy pediatric subjects. *Proceedings of the National Academy of Sciences USA, 106*, 6790-6795, doi:10.1073/pnas.0811221106

Green A.E., Munafo M.R., DeYoung C.G., Fossella J.A., Fan J., Gray J.R. (2008). Using genetic data in cognitive neuroscience: from growing pains to genuine insights. *Nat. Rev. Neurosci.* 9: 710-20.

Grady, D.L. *et al.,* (2013). DRD4 genotype predicts longevity in mouse and Human. *Journal of Neuroscience*, doi:10.1523/jneurosci.3515-12.2012

Haith, M.M., Hazan, C., & Goodman, G.S. (1988). Expectations and anticipations of dynamic visual events by 3.5-month-old babies. *Child Development* 59, 467-469.

Harpending, H. & Cochran, G. (2002). In our genes. *Proceedings of the National Academy of Sciences of the USA*, *99*, 10-12.

Isaacowitz, D.M. (2012). Mood regulation in real time: age differences in the role of looking. *Current Direction in Psychological Science* 21/4, 237-242.

Jones, L.B., Rothbart, M.K., & Posner, M.I. (2003). Development of executive attention in preschool children. *Developmental Science, 6*, 498-504, doi:10.1111/1467-7687.00307

Larsen, H., van der Zwaluw, C. S., Overbeek, G., Granic, I., Franke, B., & Engels, R.C. (2010). A variable-number-of-tandem-repeats polymorphism in the dopamine D4 receptor gene affects social adaptation of alcohol use: Investigation of a gene-environment interaction. *Psychological Science*, *21*, 1064-1068.

Lebel, C., Gee, M., Camicioli, R., Wielere, M., Martin, W., & Beaulieu, C (2012). Diffusion tensor imaging of white matter tract evolution over the lifespan. *Neuroimage* 60, 240-352.

Luria, A.R. (1973). *The Working Brain*. New York: Basic Books.

MacLeod, J.W. Lawrence, M.A. McConnell, M.M., Eskes, G.A. (2010). Appraising the ANT: Psychometric and Theoretical Considerations of the Attention Network Test. *Neuropsychology* 24/5, 637-651, doi:10.1037/a0019803

Markett, S.A., Montag, C., & Reuter, M. (2010). The association between Dopamine DRD2 polymorphisms and working memory capacity is modulated by a functional polymorphism on the nicotinic receptor gene CHRNA4. *J. of Cognitive Neuroscience* 22/9, 1944-1954.

Moffitt, T.E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R.J., Harrington, H.L., Houts, R., Poulton, R., Roberts, B.W., Ross, S., Sears, M.R., Thomson, W.M., Caspi, A. (2011). A gradient of childhood self control predicts health, wealth and public safety. Proceedings of the National Acad. of Sci. USA 108/72693-98.

NICHD Early Child Care Research Network (1993). *The NICHD Study of Early Child Care: A comprehensive longitudinal*

study of young children's lives (ERIC Document Reproduction Service No. ED3530870).

Niogi, S. and McCandliss, B.D. (2009). Individual differences in distinct components of attention are linked to anatomical variations in distinct white matter tracts. *Frontiers in Neuroanatomy*, 3:21.

Petersen, S.E. & Posner, M.I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience* 35, 71-89.

Posner, M.I. (2012). *Attention in a Social World*. New York, Oxford University Press.

Posner, M.I. & Petersen, S.E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*, 25-42.

Posner, M.I., Rothbart, M.K., Sheese, B.E., Voelker, P. (2012). Control Networks and Neuromodulators of Early Development. *Developmental Psychology* 48/3, 827-835, doi:10 1037/a0025530

Raichle, M.E. (2009). A paradigm shift in functional imaging. *Journal of Neuroscience* 29, 12729-34.

Posner, M.I., Sheese, B, Odludas, Y., Tang, Y. (2006). Analyzing and shaping neural networks of attention. *Neural Networks* 19, 1422-1429.

Reuter, M., Ott, U., Vaitl, D., Hennig, J. (2007). Impaired executive control is associated with a variation in the promoter region of the tryptophan hydroxylase 2 gene. *Journal of Cognitive Neuroscience* 19/3, 401-408, doi:10.1162/jocn.2007.19.3.401

Ronald, A. (2011). Is the child 'father of the man'? Evaluating the stability of genetic influences across development. *Developmental Science* 14:6, 1471-1478.

Rothbart, M.K. (2011). *Becoming who we are: Temperament, personality and development*. Guilford Press.

Rothbart, M.K., Sheese, B.E., Rueda, M.R., & Posner, M.I. (2011). Developing mechanisms of self regulation in early life. *Emotion Review 3/2,* 207-213.

Rueda, M.R., Fan, J., Halparin, J., Gruber, D., Lercari, L.P., McCandliss B.D. & Posner, M.I. (2004). Development of attention during childhood *Neuropsychologia,* 42:1029-1040.

Sheese, B.E., Rothbart, M.K., Voelker, P., & Posner, M.I. (2012). The dopamine receptor D4 gene 7 repeat allele interacts with parenting quality to predict Effortful Control in four-year-old children. *Child Development Research* vol. 2012, ID 863242, 6 pages, doi:10.1155/2012/863242

Sheese, B.E., Rothbart, M.K, Posner, M.I., White, L.K. & Fraundorf, S.H. (2008). Executive attention and self regulation in infancy. *Infant Behavior and Development* 31, 501-510.

Sheese, B.E., Voelker, P.M., Rothbart, M.K., & Posner, M.I. (2007). Parenting quality interacts with genetic variation in Dopamine Receptor DRD4 to influence temperament in early childhood. *Devel. & Psychopathology* 19**,** 1039-1046.

Sheth, S.A., Mian, M.K., Patel, S.R., Asaad, W.F., Williams, Z.M., Dougherty, D.D., Bush, G., & Eskander, E.N. (2012). Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature* 488, 218, doi:10.1038/nature11239

Smith, H.J., Sheikh, H.I., Dyson, M.W., Olino, T.M., Laptook, R.S., Durbin, C.E., Hayden, E.P., Singh, S.M., & Klein, D.N. (2012). Parenting and Child *DRD4* Genotype Interact to Predict Children's Early Emerging Effortful Control. *Child Development*, 83/6, 1932-1944.

Spector, T. (2012). *Identically different: why you can change your genes*. London: Orion Publishing.

Supekar, K. & Menon, V. (2012), Developmental maturation of a dynamic causal control signals in higher-order cognition: a neurocognitive network model. *PLOS Computational Biology* 8/2 e1002374,

doi:10.1371/journal.pcbi.1002374

Tang, Y-Y., Lu, Q., Fan, M., Yang, Y., & Posner, M.I. (2012). Mechanisms of White Matter Changes Induced by Meditation, *Proceedings of the National Academy of Sciences USA* 109 (26), 10570-10574, doi:10.1073/pnas.1207817109

Tang Y., Lu Q., Geng X., Stein E.A., Yang Y., & Posner M.I. (2010). Short-term mental training induces white-matter changes in the anterior cingulate, *PNAS* 107, 16649-16652.

van Ijzendoorn, M.H., & Bakermans-Kranenburg, M.J. (2006). DRD4 7-repeat polymorphism moderates the association between maternal unresolved loss or trauma and infant disorganization. *Attachment and Human Development, 8*, 291-307.

Vygotsky, L.S. (1934/1962). *Thought and language*, E. Hanfmann & G.Vakar (eds), Cambridge MA: MIT Press.

Westlye, L.T., Grydeland, H., Walhovd, K.B. & Fjell, A.M. (2011). Associations between Regional Cortical Thickness and Attentional Networks as Measured by the Attention Network Test. *Cerebral Cortex* 21/2, 345-356, doi:10.1093/cercor/bhq101

Wang, E.T., Kodama, G., Baldi, P., & Moyzis, R.K. (2006). Global landscape of recent inferred Darwinan selection for Homo sapiens. *Proceedings of the National Academy of Science, 103*, 135-140.

Wynn, K. (1994). Addition and subtraction by human infants. *Nature*, 358:749-750.

Zilles, K. (2005). Evolution of the human brain and comparative syto and receptor architecture. In S. Dehaene, J.-R. Duhamel, M.D. Hauser and G. Rizzolatti (eds), *From Monkey Brain to Human Brain*: 41-56. Cambridge Mass: MIT Press, Bradford Books.

New Interfaces for the Brain

John Donoghue*

Between Man and Machine

We interact with the world by receiving input from an exquisite group of sensors that provide the brain a filtered sample of the world and a set of artfully arranged muscles that provide the brain's only conduit for action. The brain's vast neural networks, joining billions of neurons, link percepts derived from sensors to thoughts, memories, and actions. Nearly infinitely-flexible connection patterns in the human brain permit remarkable behaviors, ranging from gymnastic leaps to piano arpeggios, to paintings, to speech. Vast sensorimotor interactions, flavored by emotion, drives, and reveries, are the essence of our humanity. How activity patterns emerge from the networks of our brain to create human activity remains one the greatest mysteries of science. Diseases or disorders that disconnect the brain from the world profoundly influence our being. Here, I will focus on the loss of the connections from the brain to the muscles and efforts to combine the knowledge from neuroscience, engineering and medicine to provide unique technology to reconnect the brain to the outside world for people with paralysis. I will then discuss the potential impact of this emerging neurotechnology on what it means to be human.

When we express a behavior, patterns of electrical impulses emitted from the cerebral cortex pass to the brainstem and spinal cord through a pencil-lead thick fiber bundle possessing about 1 million axons. It is this route by which commands from cortex reach the rest of the nervous system. This tenuous bundle of cortical fibers orchestrates voluntary movement patterns in the spinal cord, which in turn coordinates volitional muscle action. Strokes, traumatic brain injury, or degenerative diseases can damage the brain's ability to address the muscles, leading to paralysis. Damage to the corticospinal pathway anywhere along its path to the spinal cord, disconnects the brain from everything below the damage, producing voluntary actions of those body parts. In the worst case when the injury is high in the neuraxis tetraplegia results, making a person fully dependent on others because no useful body movements are possible. There is no cure for this devastating damage, which robs the person of a critical part of their humanity – the freedom to act on the world as they wish. However, these people with such disorders frequently retain their full complement of cognitive abilities, but they cannot express their thoughts. People with tetraple-

gia, as occurs after cervical spinal cord injury or brainstem stroke, or neurodegenerative diseases like amyotrophic lateral sclerosis (ALS; Lou Gerhig's disease), are completely reliant on others for even the most basic of human functions like eating or grooming. With locked-in syndrome, which is the most severe form of disconnection, people can be fully conscious but unable to speak or make nearly any useful movements, limiting communication to slow and tedious repetitions of a few spared actions, like eyeblinks. Even less debilitating paralysis from disease or injury, as might result from a smaller stroke, limits action of one part of the body, not uncommonly one arm. Loss of arm function can also severely restrict everyday self-care as well as skilled activities needed for employment or recreation. There is excitement and hope for stem cell therapy to repair damaged axons or replace lost neurons; this biological repair strategy appears to have an encouraging, but still complex and protracted path ahead before humans will benefit (Sahni and Kessler, 2010). Advances in neurotechnology and neuroscience in recent years are beginning to show that a physical repair of damaged pathways is a viable alternative path for neurorestoration.

Neurotechnology as a physical bridge from brain to the outside world

A revolution in neurotechnology is providing a new approach to restore lost functions (Donoghue, 2008). Neural interfaces are devices that couple with the brain to sense, adjust, or replace activity. That is, neural interfaces have the potential to either *write in* signals to the brain by stimulating neurons, or activity can be *read out* by sensing ongoing activity. One class of already successful *write in* neural interfaces is based on electrical stimulation that is used to influence brain circuits, an approach called neuromodulation. Deep brain stimulation (DBS) is one established success of a neuroan interface of this type. Patterns of localized electrical stimulation in the subthalamic nucleus (STN, a small collection of neurons in the center of the brain) are able to reestablish a balance of activity in Parkinson's disease (PD), restoring the ability to resume everyday functions (Weaver *et al.*, 2009). Stimulation in the STN reduces overactivity in the muscles, reducing tremor and rigidity. The exact mechanism by which disrupted circuits reestablish balance is not clearly understood but it is thought to modulate the interactions of circuits that have lost their appropriate interactions to control movement. While DBS restores the ability to move more ably, it does not slow the ongoing degeneration of dopaminergic neurons that cause PD. Implantation of DBS electrodes for neuromodulation is no longer rare: more than 80,000 people have already had these mm-size electrodes that are permanently inserted approximately 4 cm into the brain. Neuro-

modulation technology is also being tried for debilitating mood and behavioral disorders like severe depression or obsessive compulsive disorder, with encouraging initial results (Benabid and Torres, 2012). However, the ability to manipulate mood and affect by external interventions raises important ethical issues that require careful, ongoing evaluation as these technologies become more and more broadly applied (Kuhn *et al.*, 2009). Another form of *writing in* employs stimulation to replace lost senses. Patterned stimulation of the auditory nerve within the cochlea using cochlear implants has been enormously successful in restoring useful sound perception in more than 200,000 people with deafness, including many children. In addition, stimulation of the retina, by a small array of electrodes placed inside the eye of those with blindness from photoreceptor degeneration, is beginning to show promise as a way to provide at least crude visual perception (Humayun *et al.*, 2012). These devices are replacement parts for lost inputs that can restore more useful interaction with the world for people with profound sensory loss.

Reading out for action restoration

Neurotechnology to read out or sense neural signals as a direct control source to restore movement for people with paralysis is in early stage development, but showing considerable promise. A brain computer interface (BCI) is a neural interface system designed to form a new bridge from the brain to the world, bypassing injured motor pathways that prevent volitional action. Stroke, spinal cord injury, or degenerative diseases often leave cerebral motor structures intact but block the route for commands to reach the body when they damage the corticospinal pathway. Thus, a principle goal of a BCI is to read out motor intentions from their origins in the brain and deliver a derived command to assistive technologies that can restore useful functions. BCIs employ sensors to detect neurally-based movement signals, which are subsequently translated into motor commands using computer algorithms that interpret or *decode* neural activity. These commands then control devices that can carry out lost functions, such as those that are ordinarily performed by the arm. Some types of BCIs attempt to capture any recordable brain signal that can be co-opted to become a new communication channel (see: Wolpaw and Wolpaw, 2012). Evoked or brain state related potentials, such as the electroencephalogram (EEG), can be used as a 'switch' typically to indicate a yes-no like choice. EEG signals, which reflect a more global intention or brain state rather than exact movement commands are of interest because they can be obtained from the scalp. More high fidelity sensors must be placed in contact with the brain. However,

even this minimal enhancement can be highly valuable for people who are locked-in, without any useful means to express their thoughts or needs. Others are developing technology that has a more ambitious goal – to reconnect the brain's motor areas to devices that can fully perform the lost action, either through a replacement device such as a robot or computer, or through a physically reconnection between the brain to the body that could reanimate the paralyzed muscles themselves.

The BrainGate Pilot Human Clinical Trial

Initial clinical trials of BCIs to restore movement control and independence for people with profound paralysis have already begun. Our research group is developing a 'first of its kind' BCI, called BrainGate (www.braingate2.org). This neural interface system is being designed to allow people with paralysis to use their own volitional movement signals to operate a range of assistive technologies. Our major emphasis is to emulate arm function for those that cannot use their arm, because of importance of the arm in so many everyday human activities. BrainGate consists of a tiny 4 x 4 mm sensor that is implanted in the arm region of the motor cortex, where commands to move are accessible and partially understood. Neural signals recorded from even a small patch of motor cortex can reflect both intended hand movement in space and a hand squeeze. These intentions can be decoded from patterns of neural activity into machine commands sufficient to operate a computer, control robotic assistants, or potentially even drive paralyzed muscles themselves. In proof of concept preclinical studies we demonstrated that BrainGate could be safely implanted in animals and provide useful control signals (Serruya *et al.*, 2002); other work in animal models has extended this initial observation and demonstrated a wide potential to achieve complex control based on a remarkably small sample of cortical activity (e.g., Taylor *et al.*, 2002; Carmena *et al.*, 2003; Hatsopoulos *et al.*, 2004; Achtman *et al.*, 2007; Velliste *et al.*, 2008; Vargas Irwin *et al.*, 2010). These basic science successes led to testing of the BrainGate neural interface system in people with paralysis[1] (Hochberg *et al.*, 2006; 2012; Simeral *et al.*, 2011).

Sensors, Signals and Sources

The fundamental problems to be solved for a BCI are: first, to localize and detect sources of movement control signals in the brain; second, de-

[1] Caution: investigational device. Limited by federal (USA) law to investigational use.

code or translate them into useful commands, and third, to identify or develop assistive technologies that can be usefully operated directly from the brain by people with paralysis. We know from years of neuroscience research that the primary motor area (known as M1), which is essential for volitional action in humans, resides within a strip of cerebral cortex coinciding with the precentral gyrus on the surface of each cerebral hemisphere. M1 is divided into separate lateral to medial regions that command face, arm and leg, and is a major source of volitional commands that run in fiber pathways to the spinal cord (for the body) and brainstem (for the head). Neurons in the M1 arm area work together as ensembles, creating activity patterns that relate to the motion of the hand in space, or movement of the fingers and wrist. We know enough about M1 activity patterns from recordings in animal studies to estimate an intended action, such as move the arm up, left, down or right, or open and close the hand; it is likely that much more information about the arm could be harvested from neural activity patterns. Other cortical areas have arm related signals as well (Kalaska *et al.*, 1997; Sathanam *et al.*, 2006) and these areas may also provide useful command signal sources in humans, especially if M1 itself were to be damaged.

Because neurons are tiny and their signals are weak, recordings are obtained from an array of hair-thin electrodes that must be placed very near individual cells. These microelectrodes detect trains of millisecond-long impulses known as spikes (or action potentials); the spiking of many neurons (typically up to a few dozen) provides an evolving pattern in time, with each neuron emitting a pulse sequence much like a series of 0s and 1s. Device placement requires an invasive procedure by a neurosurgeon where the brain is exposed and the array placed. For the BrainGate trial the 4 x 4 mm array has 100 electrodes, each 1.5 mm long, arranged in a 10 x 10 matrix (Rousche and Normann, 1998). We adapted this sensor so that it could be permanently implanted in the human M1 arm cortex. The initial sensor for humans has a cable running from the array in the brain, out across the skull to a small plug mounted on the head that emerges through the skin. This percutaneous plug is all that is visible once surgery is completed. The plug serves as the connection to the MI implanted array for the few hours of each recording and testing session in the pilot study. Currently, neural signals are processed and decoded by a nearby rack of electronics that converts the pattern of neural signals into commands. When our participant imagines moving their arm, the resultant volitional neural activity leads to the movement of a cursor on a computer screen or other assistive technologies, such as a robot arm.

Useful actions after years of paralysis

The pilot clinical trial is testing the safety and potential utility of the BrainGate neural interface system. So far, seven people with tetraplegia have had an array implanted and have been part of our trial. All were enrolled more than one year since they had their stroke, spinal cord injury or onset of ALS (Lou Gerhig's disease; motor neuron degeneration); they had all been consented through an approved, institutionally and FDA guided process under an Investigational Device Exemption (IDE).

Our participants have shown that despite their long-standing paralysis, the motor cortex remains fully able to generate movement commands, as if the arm were actually engaged in reaching and grasping when they imagine these actions. This has been a surprising finding in that this part of the brain, thought to be key for volitional movement, had lacked a meaningful output for years. Nevertheless, yet it seemed to act as if it were commanding actual movement. Using their own decoded M1 signals participants have operated a computer to emulate mouse pointing and clicking actions (by imagining moving their hand and then squeezing to click). Neural patterns based only on thoughts of (imagined) arm and hand actions allow them to navigate software and type messages. People with severe paralysis from stroke have also been able to use robotic arms to perform reach and grasp actions. One participant unable to move or speak, recently was able to serve herself her own morning coffee for the first time in nearly 15 years (Hochberg *et al.*, 2012). These self directed, BCI-based actions are slower and less accurate than able-bodied people, but they are major advances for those completely reliant on others. It is important to note that in the study, participants only use the system for a few hours a week when they engage in a session; the technology is not yet available for everyday use. Control is not always useful. Changes in the signal can degrade performance. These are issues that only can be resolved by working with the system to advance its reliability and stability, which is now a major effort in a large number of laboratories interested in furthering BCIs for people with paralysis.

Promise of BCIs

These first human clinical tests are small steps towards restoring independence and dignity to people deprived of the ability to voluntarily perform everyday actions, freeing them from reliance on others often for even the most basic human activities. The current BCI systems remain to be made better, faster and more reliable, with implanted wireless sensors that replace the head plug and a tethering cable now required. Others are testing sensors that may make command signals more reliable, stable, or longlasting.

Promising advances in neurotechnology are underway but will still take a number of years of engineering and science, at great expense and effort, before they are generally available. While the current systems are being evaluated in severely disabled people, technology like BrainGate could help millions worldwide who have more limited paralysis, particularly from stroke which affects hundreds of thousands of people each year. A successful BCI could provide a means for them to once again move their paralyzed arm or stand and walk. Connecting brain to muscles using technology is not as remote as it might seem. Neural signals could be routed to electrical stimulators implanted in the arm that can activate muscles. Functional electrical stimulation systems have already been deployed in people with spinal cord injury, although commanded by switches rather than the brain (Peckham and Knutson, 2005). A BCI to FES system would effectively create a physical bridge to replace the missing neural link between brain and muscles and allowing volitional control over one's own body.

Beyond clinical applications, sensors used to detect neural movement activity patterns are also providing a unique glimpse into the human brain in everyday operation and a cellular level picture of the effects of disease and damage. The recordings of a very small ensemble of neurons reveal a rich and complex processing in which neurons reflect much more than a movement command, raising more questions about the processes that lead from perception and thought to action in the human brain. In the very long term this physical nervous system has a major goal to restore dignity and control by making the person indistinguishable from any other, fully capable of pursuing goals available to any fully functioning person.

Neuroethics

Restoring people to a level where they can realize all of their potential is hardly an issue that most would consider to be controversial. However, advances in understanding how the brain plans, organizes and transforms thoughts to become actions could lead to applications that might extend human capabilities. These applications present ethical challenges: if we could augment human performance by direct brain control, who would have access, who would pay, would it be appropriate, how would it be controlled? What would it mean to be human if the brain was fused to a machine? Some of these same challenges are present for neurally active pharmacological agents or even augmenting external prosthetic limbs which may serve as a model for BCIs. At present we are a long way from meaningful readout of a brain that is good enough to infer innermost mental activity. Such a feat may never be met because the full richness of human cognitive

behavior seems to emerge at a level that might be immeasurable. In addition the need for surgical procedures, which is highly regulated and monitored, presents a substantial barrier to entry for mundane or recreational applications of invasive BCIs. However, the responsible conduct of scientists, engineers, clinicians and a wide range of societal organizations in this emerging field requires that there be an ongoing debate of the use of neural interface technology and its implications for what it means to be human.

References

Achtman N, Afshar A, Santhanam G, Yu BM, Ryu SI, Shenoy KV. 2007. Free-paced high-performance brain-computer interfaces. *J Neural Eng* 4:336-47.

Benabid AL, Torres N. New targets for DBS. Parkinsonism Relat Disord. 2012 Jan; 18 Suppl 1:S21-3. Review.

Carmena JM, Lebedev MA, Crist RE, O'-Doherty JE, Santucci DM, *et al*. 2003. Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biol* 1:E42.

Christopher & Dana Reeve Foundation, 2009. *One Degree of Separation: Paralysis and Spinal Cord Injury in the United States*, Short Hills, NJ.

Donoghue JP. 2008. Bridging the brain to the world: a perspective on neural interface systems. *Neuron* 60:511-21.

Hatsopoulos N, Joshi J, O'Leary JG. 2004. Decoding continuous and discrete motor behaviors using motor and premotor cortical ensembles. *J Neurophysiol* 92:1165-74.

Hochberg, Leigh R *et al.*, 2006. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099), pp. 164-71.

Hochberg, Leigh R *et al.*, 2012. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398), pp. 372-5.

Humayun MS, Dorn JD, da Cruz L, Dagnelie G, Sahel JA, Stanga PE, Cideciyan AV, Duncan JL, Eliott D, Filley E, Ho AC, Santos A, Safran AB, Arditi A, Del Priore LV, Greenberg RJ; Argus II Study Group. Interim results from the international trial of Second Sight's visual prosthesis. *Ophthalmology*. 2012 Apr;119(4): 779-88.

Kalaska JF, Scott SH, Cisek P, Sergio LE. 1997. Cortical control of reaching movements. *Curr Opin Neurobiol* 7:849-59.

Kuhn J, Gaebel W, Klosterkoetter J, Woopen C. Deep brain stimulation as a new therapeutic approach in therapy-resistant mental disorders: ethical aspects of investigational treatment. *Eur Arch Psychiatry Clin Neurosci*. 2009 Nov; 259 Suppl 2:S135-41. Review.

Peckham PH, Knutson JS. 2005. Functional electrical stimulation for neuromuscular applications. *Annu Rev Biomed Eng* 7:327-60

Rousche PJ, Normann RA (1998) Chronic recording capability of the Utah Intracortical Electrode Array in cat sensory cortex. *J Neurosci Methods* 82:1-15.

Sahni V, Kessler JA. Stem cell therapies for spinal cord injury. *Nat Rev Neurol*. 2010 Jul;6(7):363-72.

Santhanam, G. *et al.*, 2006. A high-performance brain-computer interface. *Nature*, 442(7099), pp. 195-8.

Serruya MD, Hatsopoulos NG, Paninski L, Fellows MR, Donoghue JP. 2002. Instant neural control of a movement signal. *Nature* 416:141-2.

Simeral, J D *et al.*, 2011. Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array.

Journal of neural engineering, 8(2), p. 025027. Available at www.ncbi.nlm.nih.gov/pubmed/21436513 [Accessed July 12, 2012].

Taylor, Dawn M, Tillery, S.I.H. & Schwartz, A.B., 2002. Direct cortical control of 3D neuroprosthetic devices. *Science (New York, N.Y.)*, 296(5574), pp. 1829-32. Available at www.ncbi.nlm.nih.gov/pubmed/12052948

Vargas-Irwin CE, Shakhnarovich G, Yadollahpour P, Mislow JM, Black MJ, Donoghue JP. Decoding complete reach and grasp actions from local primary motor cortex populations. *J Neurosci*. 2010 Jul 21;30(29):9659-69.

Velliste, M. *et al.*, 2008. Cortical control of a prosthetic arm for self-feeding. *Nature*, 453(7198), pp. 1098-101. Available at www.ncbi.nlm.nih.gov/pubmed/18509337 [Accessed July 13, 2012].

Weaver FMP, Follett KMDP, Stern MMD, *et al*: Bilateral deep brain stimulation vs. best medical therapy for patients with advanced Parkinson disease: a randomized, controlled trial. *JAMA* 2009; 301:63-73.

Wolpaw, J. & Wolpaw, E., 2012. *Brain-Computer Interfaces: Principles and Practice*, 1st ed., Oxford University Press, USA.

Zhuang, J. *et al.*, 2010. Decoding 3-D reach and grasp kinematics from high-frequency local field potentials in primate primary motor cortex. *IEEE transactions on biomedical engineering*, 57(7), pp. 1774-84.

Homo Docens and the Teaching Brain

Antonio M. Battro

A Socratic Dialog revisited

> ἔχεις μοι εἰπεῖν, ὦ Σώκρατες, ἆρα διδακτὸν ἡ ἀρετή; ἢ οὐ διδακτὸν ἀλλ᾽ ἀσκητόν; ἢ οὔτε ἀσκητὸν οὔτε μαθητόν, ἀλλὰ φύσει παραγίγνεται τοῖς ἀνθρώποις ἢ ἄλλῳ τινὶ τρόπῳ;

Plato founded his famous Academy in Athens two thousand and four hundred years ago where many of the brightest minds of Greece were educated. They were "doing philosophy together", *sumphilosophein* (Berti, 2010). Socrates, the great teacher, was the master figure in the Platonic dialogs. He excelled in the way he presented the questions and negotiated the answers but, paradoxically, Socrates himself, the master teacher, tried to show that *he was not teaching at all, he was just helping the others to unfold their own knowledge*. Today we are "doing science together" even if we are separated by great distances. Teamwork in the digital era has distance 0 and no borders. In this context we can ask two questions: could we use a Socratic dialog today as an effective pedagogical tool with large populations of students online? Could we identify the brain processes of the teacher and the student involved in such a quest? The following considerations will hopefully suggest some answers.

Plato in *Meno*, a dialog about virtue, showed one of the examples of Socrates' peculiar way of "teaching" in great detail (see the *Meno* dialog in www.perseus.tufts.edu). Meno asked Socrates "whether virtue is acquired by teaching or practice; or if neither by teaching nor practice, then whether it comes to man by nature, or in what other way". In the search for an answer to this crucial question Socrates presented a detailed proof of his peculiar theory of teaching by giving a "lesson" on geometry to an ignorant boy who was Meno's slave: the problem was to double the size of a given square.

"Attend now to the questions which I ask him, and observe whether he learns from me or only remembers," said Socrates to his friend Meno when he started his lesson about the duplication of the area of a square. But Socrates also wanted to prove a cognitive thesis, a most controversial one: "Do you observe, Meno, that I am not teaching the boy anything, but only asking him questions..." And he insisted that he was not teaching at all: "Mark now the farther development. I shall only ask him and not teach him, and he shall share the enquiry with me: and do you watch and see if

you find me telling or explaining anything to him, instead of eliciting his opinion". The dialog ended with the solution of the geometrical question by the slave. And then Socrates asked Meno: "What do you say of him, Meno? Were not all these answers given out of his own head?" "Yes – answered Meno – they were all his own". This lesson became a classical paradigm of Socratic pedagogy for centuries. In fact, it is perhaps the first time in history that somebody recorded in detail all the questions and answers of an exchange between teacher and pupil on a very precise topic. I think that this is one of the most beautiful pieces of pedagogy ever done.

Seven steps in search of the teaching brain

I will now follow several steps of the long path we have taken to understand this particular model of teaching. The question is: can we really teach how to think*?* (Battro, 1977).

1) Inspired by Meno

I became interested in *Meno* since the very beginning of the introduction of computers in education, in the early eighties, using *Logo* as a digital tool (Papert, 1981; Battro, 1998). In fact, most of the answers of the "slave/student" dialog were given by yes or no, a very convenient outcome to analyze with the help of a computer. Thirty years later our expectations were fulfilled (see the Sixth step).

2) The teaching brain

Soon we became aware that we were still lacking substantial support from the neurocognitive sciences to explain why all animals learn but only humans are so performing in the difficult art of teaching, even since early childhood (Strauss, 2005; Battro, 2010). Animals cannot teach in the way humans do (Caro & Hauser, 1992; Passingham, 2008), but until now most of the scientific effort in neuroeducation (Battro, Fischer & Léna, 2008) has been focused on the neural basis of learning, on the *learning brain* but not on the *teaching brain* (Battro, 2010). Fortunately we can correct this serious bias today using brain-imaging technologies in an experimental classroom setting.

3) Neuropedagogy

In order to establish a new field of knowledge we should take a "transdisciplinary attitude" (Koizumi, 2008). This is what happens in many laboratories of cognitive neurosciences today (see: IMBES, International Mind, Brain and Education Society; www.imbes.org). We may call *neuropedagogy* this new field of the theory and practice of teaching and learning.

4) A standardized Socratic dialog

At the Laboratory of Integrative Neurosciences of the University of Buenos Aires directed by Mariano Sigman we decided to explore *Meno* as a classical "lesson" of relevant significance in the story of Western pedagogy. The *Meno* dialog was translated into a standard format of 50 questions (in Spanish) that repeated the Socratic master class with 58 high school and college students, in "one to one" teacher/student interactions using pen and paper to draw the figures (Goldin, Pezzatti, Battro & Sigman, 2011).

We first parsed the dialog in linear and conditional branches. Conditional branches diverge from questions in which the slave makes an error and are only transited if the participant makes exactly the same error. For instance, in Question 10, Socrates asks "This (side) is two feet long: what will be the side of the other (square) which is double in size?" Meno's slave responded: "Clearly, Socrates, double" which is an error because the new square would be four times the size of the given one.

The experimental results show a remarkable agreement between Socratic and empiric dialogs: "In 28 questions, the response of every single participant followed precisely the Socratic dialog, as Meno's slave did some two thousand four hundred years ago! In questions in which Meno's slave made a mistake, within an unbounded number of possible erred responses, the vast majority of empiric responses coincided with the error of the dialog" (Goldin *et al.*, 2011).

5) Comparative studies

In order to compare these results with other cultures we asked our colleague Jiaxian Zhou of the Center for Educational Neuroscience, East China Normal University at Shanghai, to repeat the Socratic dialog with her Chinese students using our methodology. The results of the students of Buenos Aires and Shanghai were similar (Jiaxian Zhou, personal communication). This finding suggests that the cognitive process involved in the solution of the Socratic problem may be common to students of different cultures, but clearly we would need larger populations to show the "universality" of this neurocognitive process.

6) Meno online

We can now use the digital platform provided by OLPC, One Laptop Per Child www.laptop.org (Negroponte, 2007), to reach a large number of students online by using a software that automatically asks the Socratic questions to the student and can follow, guide and track the answers step by step in a digital version of the *Meno* dialog. By the way, this computer-based lesson on geom-

etry "without teacher" could be understood as a metaphor of the paradoxical Socratic statement that a teacher doesn't teach (nor does a computer...).

7) Transfer of knowledge

It was a complete surprise for us to discover that almost 50% of the participants that reached the correct solution: "take the diagonal of the given square as the side of the square with a doubled area" *failed* when asked to double the area of a square of a *different* size! They couldn't *generalize* the geometrical construction to a square of any size… This transfer failure is amazing and suggests that the Socratic dialog per se is not enough to induce a student to overcome a cognitive bias (such as to double the length of the side in order to double the size of the square) and accept a stable and definitive solution that can be generalized to any square.

This typical regression to a former and erroneous state of knowledge clearly shows that the lesson wasn't understood and the new knowledge wasn't assimilated. This phenomenon has been recognized in many disciplines, for example in the teaching of physics where advanced students must "unlearn" the Aristotelian notion of force to simulate a landing on the moon using a computer program (di Sessa, 1981). We can conclude therefore that the Socratic method per se, as shown in the *Meno* dialog, has a low pedagogical efficiency. This unexpected failure in the process of generalization is a major educational problem that needs further clarification. We decided then to explore this issue with the tools of the new neuropedagogy.

8) A dialog between brains

We started a new research using portable brain imaging equipment (functional near-infrared spectroscopy, fNIRS) on both teacher and student during the *Meno* dialog. As a control after the Meno experiment the teacher and the student read aloud a passage of Plato's dialog *Apology of Socrates*, taking respectively the role of Socrates and Meleto. The main result with *Meno* is that the left prefrontal area recorded in this experiment with 17 teacher/student couples showed significant differences of cortical activation between the students that could generalize the solution to any square in comparison to those that didn't generalize at all. The former showed a minimum level of activation – similar to the teacher's brain pattern – while the latter maintained a higher activation. Also, students that would later show a sound transfer of knowledge showed a drop in activity during the whole dialog while students that could not generalize showed sustained levels of activity during the entire dialog (Holper, Goldin, Shalóm, Battro, Wolf & Sigman, submitted to publication).

Another interesting finding is that specific brain activations during a given lesson *may predict the success of the learning process of the student*, in other words *we may predict the level of efficiency* of the learning processes of an individual, in this case if the student can generalize or transfer the knowledge just acquired to another situation. In *Meno* the level of brain activity of the teacher serves as a *standard of efficiency* for a given task. We absolutely need the records of both the learning brain and the teaching brain to understand the cognitive process involved during a lesson. Students that transfer show a positive correlation with the levels of cortical activity of the teacher, while a negative correlation occurred with the group of students that couldn't transfer.

After the middle of the dialog, Socrates shifted the focus of the arguments to the diagonal of the square and, with question 33, launched a new path in the geometrical reasoning that will end in the solution of the problem. We call that key question the *diagonal argument*. At that moment we observed a discontinuity revealing a small effect of student transfer in the teacher's NIRS signal. It seems that in such successful interventions student and teacher brains "dance at the same pace", but we need more evidence to affirm this.

A possible interpretation is that the geometric solution is correctly assimilated by the student only when the brain has reached a higher level of *efficiency* for this particular task, in other words "doing more with less" neural activity. This increase in efficiency is what we see in the reduced level of brain activation of the teacher during the dialog. Instead, when the student is unable to generalize, he or she still needs to sustain a higher level of neural activation. In this sense we agree with the recent statement of Bullmore and Sporns (2012) that "*the brain is expensive*, incurring high material and metabolic costs for its size – relative to the size of the body – and many aspects of brain network organization can be mostly explained by a parsimonious drive to minimize these costs". We believe that these costs can be reduced by a sound (neuro) pedagogy that enhances the efficiency of the neural networks in place. Of course our results are only a first and very modest step in the long way to understand what good teaching is.

In conclusion, we can expect that in the near future low-cost and high-performing wireless and portable brain imaging equipment will be common in experimental classroom settings and will, hopefully, help to transform the way we teach and learn. A whole new world will be then opened to education.

Acknowledgments

I wish to express my gratitude to my colleagues of our international team interested in the study of the "teaching brain", in particular to:

– Mariano Sigman, Andrea Goldin, Laura Pezzatti, Diego H. Shalom, Diego Fernández Slezak, and Matías López, *Laboratorio de Neurociencias Integrativas. Facultad de Ciencias Exactas, Físicas y Naturales, Universidad de Buenos Aires.*

– Lisa Holper and Martin Wolf, *Biomedical Optics Research Laboratory (BORL), University Hospital, Zurich.*

– Jiaxian Zhou, *Center for Educational Neuroscience, School of Psychology and Cognitive Science, East China Normal University, Shanghai.*

References

Battro, A.M. (1978). ¿Se puede enseñar a pensar? *Criterio*, 5, 1801, 748-752.

Battro, A.M. (1998). Intellectual Prostheses. Theory and practice. In Pontifical Academy of Sciences, *Human Genome, Alternative Energy Sources for Developing Countries, Fundamental Principles of Mathematics and Artificial Intelligence*, Proceedings of the Plenary Session 25-28 October 1994, *Scripta Varia* 92, Vatican City, online at www.pas.va/content/accademia/en/publications/scriptavaria/humangenome.html

Battro, A.M. (2007). *Homo educabilis*: A neurocognitive approach. In Pontifical Academy of Sciences, *What is our real knowledge of the human being?* Proceedings of the Working Group 4-6 May 2006, *Scripta Varia* 109, M. Sánchez Sorondo (Ed), Vatican City, online at www.pas.va/content/accademia/en/publications/scriptavaria/humanbeing.html

Battro, A.M. Fischer, K.W. & Léna, P.J. (Eds). (2008). *The Educated Brain: Essays in Neuroeducation*, *Scripta Varia* 107, Pontifical Academy of Sciences & Cambridge University Press: Cambridge.

Berti, E. (2010). *Sumphilosophein: La vita nell'Accademia di Platone.* Laterza: Roma.

Bullmore, E. & Sporns, O. (2012). The economy of brain network organization. *Nature Reviews Neuroscience* 13, 336-349.

Caro, T.M and Hauser, M.D. (1992). Is there teaching in nonhuman animals? Q. *Rev. Biol.* 67, 151-174.

del Re, G. (Ed) (1992). *Brain Research and the Mind-Body Problem. Epistemological and Metaphysical Issues.* Proceedings of the Round Table of 25 Oct 1988. *Scripta Varia* 79. Chairman: Carlos Chagas. Pontifical Academy of Sciences, Vatican City, online at www.pas.va/content/accademia/en/publications/scriptavaria/brainresearch.html

di Sessa, A. (1981). Unlearning Aristotelian physics. A study of knowledge based learning. *Cognitive Science*, 6, 37-75.

Goldin, A., Pezzatti, L., Battro, A.M. & Sigman, M. (2011). From ancient Greece to modern education: Universality and lack of generalization of the Socratic Dialogue. *Mind, Brain, and Education*, 5, 4, 180-185.

Holper, L., Goldin, A., Shalom, D., Battro, A., Wolf, M. & Sigman, M. (2013) The teaching and the learning brain: A cortical hemodynamic marker of teacher-student interactions in the Socratic dialog. *International Journal of Educational Research* 59, 110.

Koizumi, H. (2008). Developing the brain: A functional imaging approach to learning and educational practices. In Battro, A.M. Fischer, K.W. & Léna, P.J (Eds). *The Educated Brain: Essays in Neuroeducation*. *Scripta Varia* 107, Pontifical Academy of Sciences & Cambridge University Press: Cambridge.

Negroponte, N. (2007). The $100 laptop. In *Globalization and Education*, Proceedings of the PAS and PASS Joint Workshop, 16-17 November 2005, *Extra Series* 28, M. Sánchez Sorondo, E. Malinvaud & P. Léna, (Eds), Pontifical Academy of Sciences and Pontifical Academy of Social Sciences & Walter de Gruyter: New York, online at www.pas.va/content/accademia/en/publications/extraseries/globalization.html

Papert, S. (1981). *Brain storms*. Academic Press; New York.

Passingham, R. (2008). *What is special about the human brain?* Oxford University Press: Oxford.

Salviucci, P. (Ed) (1964). *Brain and Conscious Experience.* Proceedings of the Study Week of 28 Sept-3 Oct 1964, *Scripta Varia* 30. Chairman: John C. Eccles. Pontifical Academy of Sciences, Vatican City. Online at www.pas.va/content/accademia/en/publications/scriptavaria/brainandconsciousexperience.html

Strauss, S. (2005). Teaching as a natural cognitive ability: Implications for classroom practice and teacher education. In D. Pillemer and S. White (Eds). *Developmental Psychology and Social Change.* Cambridge University: Cambridge.

Statement

Neurosciences and the Human Person: New Perspectives on Human Activities Final Statement

Georges M.M. Card. Cottier, Silvia Arber, Antonio M. Battro, Timothy Behrens, Enrico Berti, Olaf Blanke, Thierry Boon Falleur, Yves Coppens, Stanislas Dehaene, Christopher D. Frith, Uta Frith, Earl K. Miller, Jürgen Mittelstrass, Martin Nowak, Svante Pääbo, H.E. Msgr. Marcelo Sánchez Sorondo, Wolf J. Singer, Nora D. Volkow

The working group met for three days to discuss questions at the boundaries of neurosciences and philosophy, with an emphasis on areas where the scientific approach is making progress, and which lie at the core of what it means to be a human person: the evolution of the human brain, the mechanisms of consciousness, the capacity for evaluation and decision making and self-control, the formation of beliefs in a social group, the sense of self, and the importance of education for human brain development. For each of these topics, we summarize here the essential conclusions and the potential points of convergence between the scientific and philosophical approaches, without denying that many of these points remained heavily debated.

Human brain evolution

Paleontological evidence, exploring the consequences of climate change, nutrition and human migrations, together with genetic evidence, pointing to a limited number of recent mutations unique to the human lineage, are shedding new light on the origins of *Homo sapiens*. Human and non-human primates share brain mechanisms, both at the level of individual circuits and areas and in the manner in which these spatially distributed systems interact and are bound together by synchronized oscillatory mechanisms to form global brain-scale assemblies. The latter mechanisms play a prominent role in processes of computation, arousal, attention, and conscious perception.

The increasing complexity of the human brain led to the emergence of novel cognitive and executive abilities that enabled *Homo sapiens* to engage in cultural evolution. Although rudimentary forms of culture, including intergeneration transmission of ways of cracking nuts, etc, have been demonstrated in apes, essential steps in this process were the conception of tools,

the growing awareness of a finite life span, the development of a symbolic communication system, the transgenerational transmission of acquired knowledge by education, the creation of social belief and value systems, social cooperation, and the concretization of mental representation systems in rituals, artistic endeavors, and social institutions.

This paleontological research, although still developing, represents a major progress over the Medieval vision of the brain, which St Thomas summarizes as follows: 'For man needs the largest brain as compared to the body; both for his greater freedom of action in the interior powers required for the intellectual operations; and in order that the low temperature of the brain may temper the heat of the heart, which has to be considerable in man for him to be able to stand erect' (Thomas Aquinas, *S. Th.* I, 91, 3 ad 1). Already Plato had the intuition that the dimension that makes humans distinct from animals and plants, addressed as mental or spiritual, is a consequence of the evolution of the brain and the ensuing cultural constructs (cf. *Timaeus*, 90 a-b).

Consciousness

Elementary mechanisms of consciousness are now increasingly being analyzed at the brain level. The events that human observers report as conscious differ from those that they cannot consciously report in multiple objective neurophysiological parameters. The sense of ownership of our body and actions, and the feeling of the first-person perspective (or self-perspective) and interoception, can be manipulated in the laboratory, and their neural correlates are being discovered using brain imaging.

The sense of the unity of consciousness, although partially illusory, may be understood as a process of convergence of the various cognitive abilities or "modules" into a single large-scale brain network or "workspace". This integration, made possible by the existence of long- and short-distance anatomical and functional connections linking higher-level brain areas, is thought to enable an internal synthesis of the innate and acquired dispositions of the individual and of his recognition of himself and his body in the world.

The practical implications of these findings are important in medical practice, in order to better understand the loss of consciousness during anesthesia or following brain lesions, to facilitate the detection of residual consciousness in locked-in patients, and to search for mechanisms underlying the disruption of the unity of consciousness in psychiatric illnesses such as schizophrenia.

The discovery that consciousness can be related to specific brain systems should not conceal the fact that most of our brain operates non-consciously. Neuroscience and psychology have discovered that many of the

brain processes for knowledge, value, decision, belief formation and social representation reside at a deep and unconscious level that remains unavailable to accurate introspection and conscious recollection. Nevertheless, through self-observation, we develop some degree of explicit self-knowledge, as well as explicit theories on how our mind and the minds of others work (theory of mind).

Values and decisions

How human and non-human primates attribute values and take decisions can also be related to a set of brain areas that provide a prospective evaluation of the consequences of actions at multiple levels. Some of these areas, which can evaluate purely imaginary and novel decisions, are the brain areas most evolved in humans compared to other primates, suggesting an expanded functionality of the valuation and decision system during the evolution of hominins. The capacity for decision making exhibits an inter-individual variability which can be correlated with neurotransmitter concentrations and with functional connectivity between brain regions and networks. At one extreme, addicted individuals show a massive disequilibrium in dopaminergic networks, which biases their entire decision system towards the pursuit of drug taking while undermining their ability for self-control. Understanding of these circuits opens up the possibility of a future treatment by restoring the operation of the valuation system and strengthening that of self-control.

Beliefs and socialization

The formation of belief systems is another cognitive domain of exceptional expansion in the human species. Perceptions and beliefs are thought to jointly arise from hierarchical brain networks that confront internal models of the world with external signals and use the corresponding error signals as a corrective mechanism. Schizophrenia is a mental illness which can be analyzed as the abnormal operation of this error-propagation device.

Although hierarchical perception and belief systems exist in monkeys and apes, they reach their full development in humans in whom an additional level of sharing attention and social information plays an essential part. The circuits of the human social brain, which have only recently started to be explored, may shed some light on how we generate an interpretation of our own self, our behavior, and our sense of responsibility and accountability. The strong feeling of belonging to a social group arises extremely early during development: even infants already express preferences for others who speak the same language. The human sense of belonging to groups

generates powerful social tendencies towards in-group cooperation and out-group exclusion.

How altruistic cooperation could have evolved is the subject of several mathematical models. They suggest that cooperation is a genuine force in evolution that may even be necessary for the emergence of any complexity in life, from pluricellular organisms to insect societies and human language. In these models, trust, generosity and forgiveness enter explicitly as evolved traits that stabilize group cooperation. Humans may be distinguished by a specific form of cooperation, 'indirect reciprocity', which relies on language to extend cooperation to new individuals based solely on their social reputation. In this respect, the cognitive sciences strengthen in a new and genuine way the philosophical notions that are at the foundation of ethical and political systems, according to which a human being is essentially a 'social animal' (Aristotle, *Politics*, I, 9, 1253 a 2).

The fundamental importance of education

Even if they are not entirely accurate, human mental representations develop continuously and can be improved by education, an activity that may well be unique to humans. Brain neuroplasticity is the mechanism by which new memories and learning occur in the brain. In the human brain it allows us to not only transmit tradition and knowledge through education, but also to shape and form personality traits. Education even permits fighting against some of the now maladaptive traits that the brain inherited from its evolution. Even a few weeks of training can modify the brain networks for attention and self-control, thus enhancing willful action over automatic reaction. Thus, any human has in himself an active capability that allows him to progress 'towards himself and his own perfection' (Aristotle, *De Anima*, II, 5, 417 b 3). Hence the importance of educating and 'training' perception, knowledge, reasoning and action, in order to attain truth, good and justice.

Developing a better understanding of how knowledge, action and decision circuits diversify to enlarge the repertoire of our species, as well as studying how the teacher's brain operates to transmit education, are essential goals for future research. The teacher's role in Plato and Thomas Aquinas is to be the instrument that helps their students not only to lead out (*educere*) their own abilities, but to also develop knowledge for themselves. There is a possible point of convergence here with theories of brain development that attribute, even to very young infants, a vast repertoire of knowledge (about objects, space, time, numbers, language...) and the ability to learn by selecting the most pertinent of these internal representations. Aristotle, somewhat similarly, identifies a distinction between potency and act: knowl-

edge pre-exists in the learner in an active way, not passive as it is in general in matter. Otherwise, the human being would not be able to acquire knowledge by himself. Therefore, 'when something pre-exists in the subject in active completed potency, the external agent acts only by helping the internal agent' (Thomas Aquinas, *De Magistro – De Veritate*, 11, 1).

Interdisciplinary convergence and its difficulties

One of the most complex questions in the interdisciplinary approach is to clarify the often different meanings that a word can have when used by the different disciplines. For instance, the analysis of the 'self' is a privileged subject for interactions between neuroscience, psychology and philosophy, reflecting the different traditions and competences of these disciplines. However, they differ with respect to the epistemological status accorded to the object of investigation, the language used for descriptions and the applied methods of research. Kant in this context distinguishes between a determining self (thought) and a determinable self (the thinking subject). The neural sciences have as their object the material aspects of the brain and the cognitive and executive functions that depend on the brain, and aim to explain various aspects of thought on this material and computational basis. Metaphysical anthropology, however, draws attention to aspects of the subject that, it suggests, may not only be thought of without matter but can also be without matter. For example, based on activities such as the perception of time – dissociated from the characteristic succession of physical movement and associated to the mental principles typical of human praxis – and the insatiable thirst for knowledge, life and happiness, realist philosophy (past and present) considers that knowing the truth and tending towards good and justice are specific to human beings. The human person, through knowledge and will, draws on the absolute and does not stop at material realities but aspires to symbolic understanding, science and perfect knowledge, and desires non-market goods, that is, goods without a price, such as dignity, mutual esteem, and happiness. In the future, examining to what extent these non-material aspirations can be linked to the material reality of the brain does not appear to be an inaccessible goal and, on the contrary, should become an important object of study for cognitive neuroscience.

Science has confirmed the existence of trillions of connections between the billions of neurons and neuronal circuits that make up the human brain, and their ramifications inside the body. Nevertheless, in general philosophers of the Socratic tradition do not agree that this leads to the conclusion that human intelligence and will are just neural events that happen in the brain.

For neuroscientists, the brain integrates all bodily functions. From the point of view of the philosophers at the meeting, this does not mean that it gives the body its ontological vital unity, which is given by the soul: '*vivere viventibus est esse*' (Aristotle, *De Anima*, II, 4, 415 b 12). For Thomas Aquinas (and contemporary thinkers of his school), this emergence or independence in acting reveals the independence of being. The being (*esse, actus essendi*) does not belong to the compound but to the intellective soul proper (the soul subsists in its esse, which it communicates to the body and takes back when the body dies and ceases to 'exist'). The being (to be more precise, *esse* as *actus essendi*) adheres immediately and thus inseparably to the subsistent form. Consequently, the human soul is thought to be incorruptible and thus immortal, created directly and individually by God.

This philosophical conception, especially the central question of the relationship between the brain and the soul, generated intense debates amongst the scientists and philosophers participating in the working group. It was pointed out by the philosophers that brain functions alone may not be sufficient to serve as a basis for ethical and ontological statements about the status of the human person. Humans with severe impairments of brain functions cannot be denied humanity and dignity. Therefore, although the scientists and the philosophers agreed on the fact that the brain gives vital unity, the philosophers' stance was that the soul is the principle differentiating between living beings and is the unifying essence. Whereas organs, the brain included, and the potencies (i.e. the intellect, the will, the senses) are called secondary principles of unity, of coordination and of operation, the individual is the first principle of action and attribution. In the perspective of the neuroscientists present at the meeting, autonomous action and self-attribution could arise solely from the spontaneous patterns of brain activity that auto-organize to provide internal models and motivations to act, including moral operations (behaviours and emotions). In the perspective of philosophers present at the meeting, autonomous action and auto-organisation is the characteristic of living beings (Aristotle, *De Anima*, 412 a 12), and many of them, like microorganisms and plants, do not have a brain but a substantial principle of unity which is the soul. So the soul is the subject but in an active and coordinative sense in living beings of the different species which becomes – in the human person – a principle and responsible subject capable of reflecting on himself/herself. Both perspectives, however, agree that "the brain acts as the neural central driving force of existence" and that "brain death is the death of the individual", as stated in the Pontifical Academy's Statement *Why the Concept of Death is Valid as a Definition of Brain Death* (2008).

Conclusions

In conclusion, the current knowledge of the organization of the human brain and how it gives rise to mental states already provides an important contribution to the issue of what the human person is. Yet like any scientific enterprise, the answers that it provides remain limited. Scientists and philosophers need to search for a better language that may bridge the gaps between the disciplines and levels of analysis. This includes the language of values, responsibility, dignity and justice. The reconstructions of the concepts of consciousness and self-consciousness, mind and soul, form and information, may help to bring together the natural sciences, the social sciences and the humanities.

Thanks to the discovery of the centrality of the brain, made by the neurosciences, we now have a new starting point for our recognition of the status of the human being. Today, we can be both actors and spectators of our own actions and of ourselves – the first-person perspective of the subjective self is complemented by the third-person perspective of neuroscience. Only a human being is capable of creating such circularity by observing the functioning of his brain from the outside with increasingly powerful instruments, while also interpreting these data from inside, based on conscious self-reflection. The consequences of this dual approach are only beginning to be explored.

In addition to contributing to this conceptual search, cognitive neuroscientists also have an important present responsibility with respect to the many challenges raised by the contemporary world. New interfaces will soon link the human brain to computers and robots, alleviating paralysis but also raising difficult ethical issues. The legal system may benefit from, but also be deeply challenged by, our improved understanding of conscious and non-conscious determinants of human behavior. Many existing human institutions, such as the prison system, may ultimately require extensive reconsideration in light of our growing understanding of the human brain and the possibility of changing and educating it. Prison (deprivation of freedom to move) should never be just a punitive institution but also, and above all else, it should protect society against dangerous individuals, act as a deterrent, and be corrective and educational for those who are imprisoned.

Tables

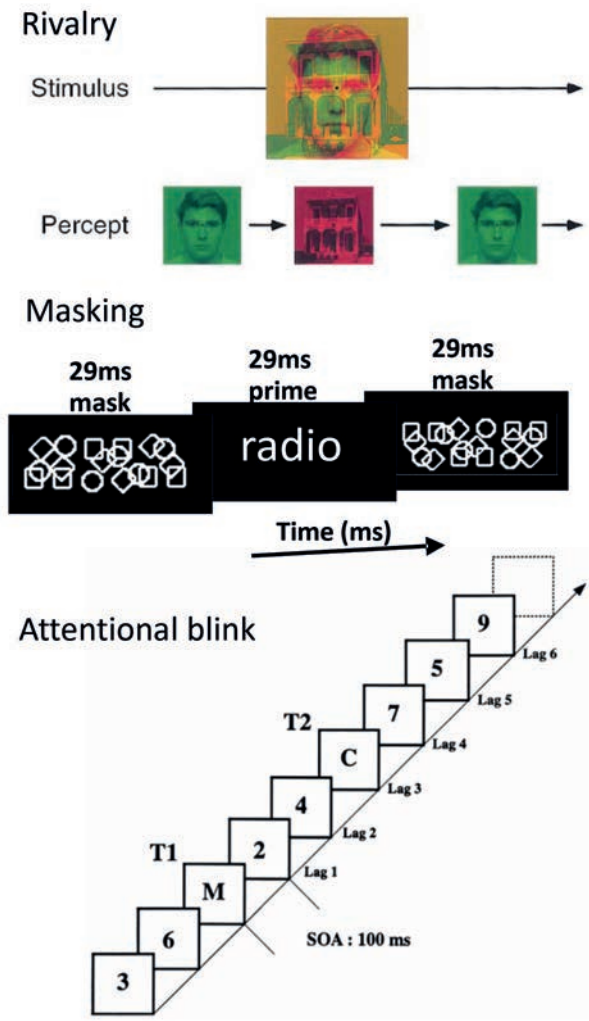

Figure 1. Examples of experimental paradigms to manipulate conscious perception. In rivarly, distinct images are presented to the two eyes, yet subjective perception alternates between seeing one and seeing the other. In masking, a visible word is made invisible by surrounding in time it with shapes that mask it. In the attentional blink, processing of a first target T1 prevents the perception of a second target T2.

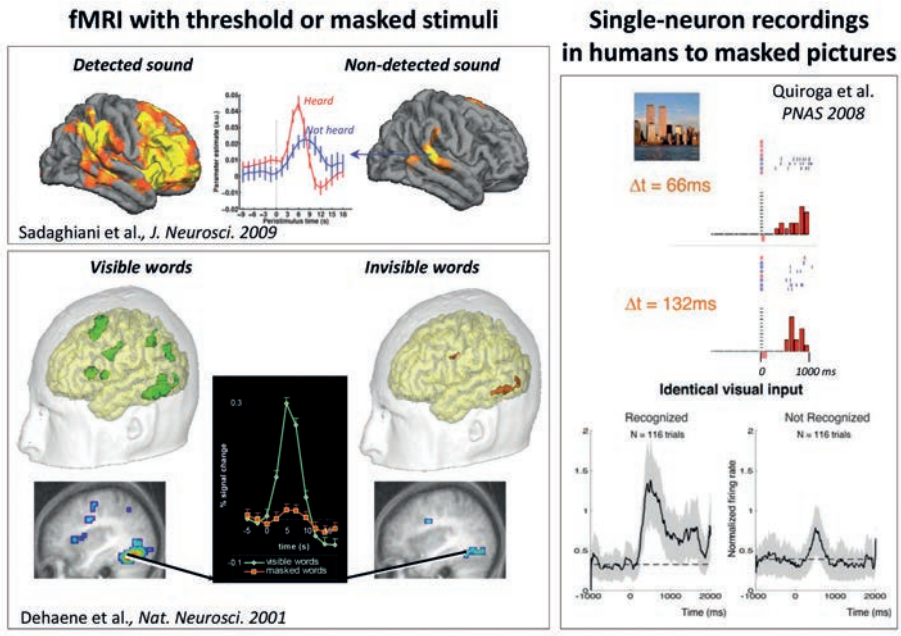

Figure 2. Converging evidence for cerebral signatures of consciousness. Conscious perception, compared to non-conscious processing, systematically involves a late and long-lasting "ignition": sensory activation is amplified and expands into a broad set of associative areas of the prefrontal, parietal and temporal lobes.

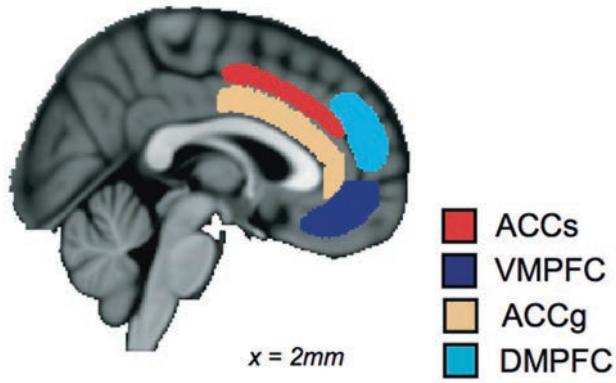

Figure 1. Medial frontal regions in a human brain. Saggital or medial view of the human brian. The frontal most part of the brain is on the right. ACCs – Anterior Cingulate Sulcus; VMPFC – ventromedial prefrontal cortex; ACCg – Anterior Cingulate gyrus; DMPFC – dorsomedial prefrontal cortex. Adapted from Behrens et al. Science 2009. For the purposes of the current review, I will not discuss the interesting differences between the sulcal (ACCs) and gyral (ACCg) portions of the anterior cingulate cortex. These are discussed at length in (Behrens et al., 2009; Behrens et al., 2008; Rudebeck et al., 2006).

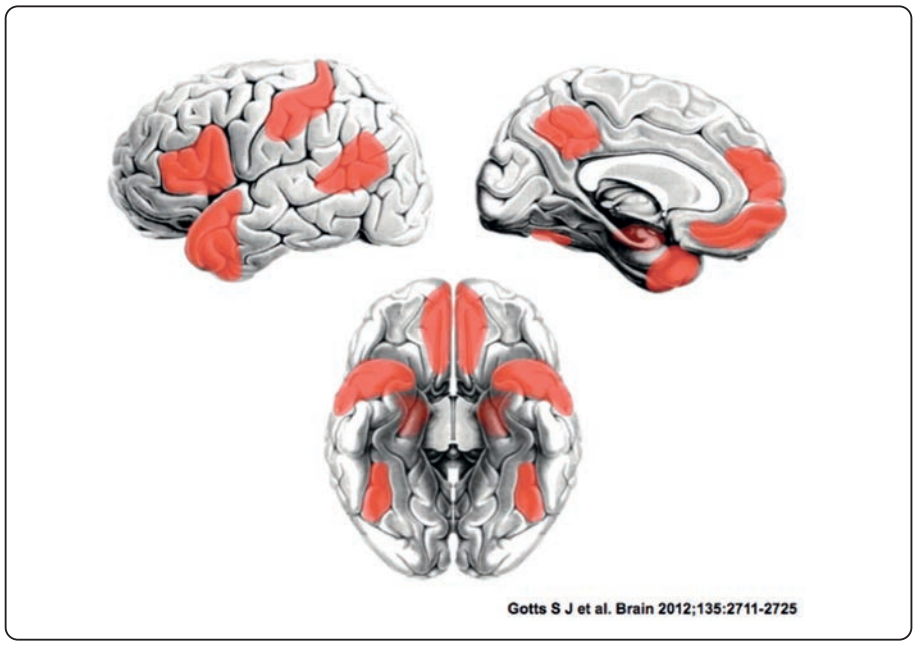

Figure 2. The social brain as revealed by autism.

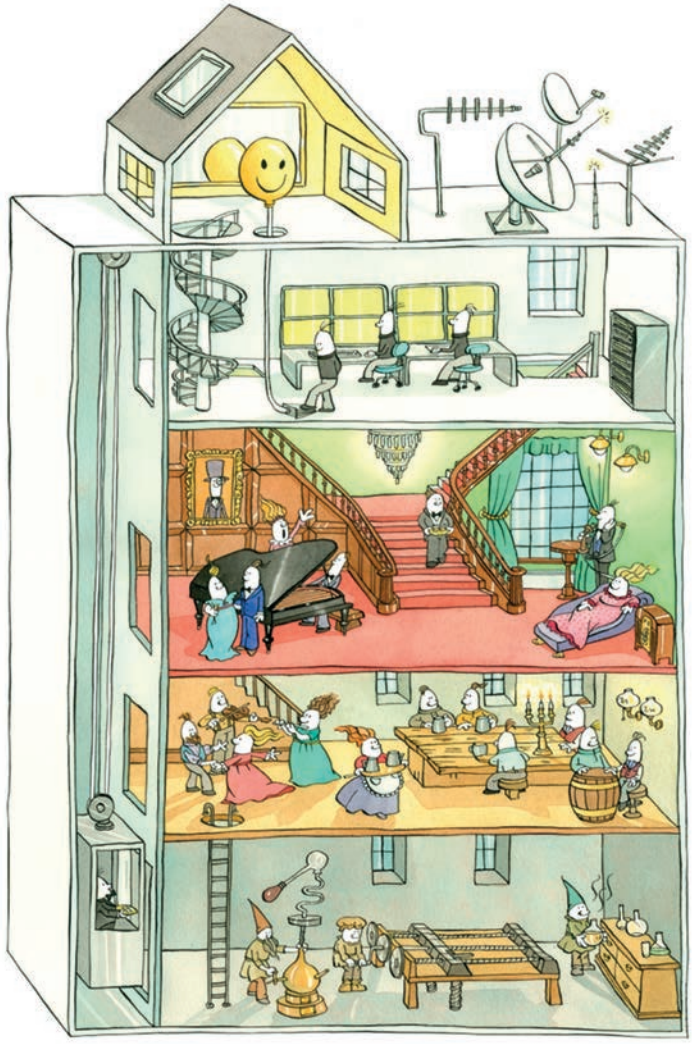

Figure 3. ©Jan McCafferty.