EMPATHY AND THE ORIGINS OF THE SOCIAL BRAIN

Allan Young
McGill University (Montreal)

VERSIONS OF HUMAN NATURE

> The topic this afternoon is…human nature in the age of biotechnology. The subject crops up … in our conversations and is very often just below the surface … [and] it's worth our while to … think about how to think about human nature in an age of genomics, in an age of neuroscience and what might be possible in the way of altering it and ultimately what those alterations might mean and whether they would be a good thing.…
>
> > Leon Kass, chairman of the session on Human Nature and Its Future, convened on 6 March 2003 by the (USA) President's Council on Bioethics

The term "human nature" commonly refers to a bundle of innate and universal human faculties and dispositions that distinguish humans from other animals and normal people from various kinds of abnormal people. Within Western societies, opinions about the contents of human nature are not uniform: they have changed over time and they vary among groups within populations. Until recently, it was permissible to speak of a canonical version of human nature, represented in the operations of key social institutions and sectors of knowledge production, notably the social and behavioral sciences, biomedicine, psychiatry, and the law.

The canonical version originated in Enlightenment debates about the nature of the mind and its relation to earlier mind-like conceptions such as the soul (Ryle 1949: 22-23). For convenience, I will refer to this version as "Human Nature 1.0." In its most basic form, Human Nature 1.0 is associated with four features:

1. Mind is the body's command-center, theater of self-awareness, and agency of self-identity and continuity.
2. Normal people are rational.

> 3. Normal people are self-interested: seeking pleasure and gratification and avoiding pain and distress.
> 4. Minds are self-contained. “[M]ental happenings occur in insulated fields, known as ‘minds’, and there is, apart from telepathy, no direct causal connection between what happens in one mind and what happens in another. Only through the medium of the public physical world can the mind of one person make a difference to the mind of another. The mind is its own place and in his inner life each of us lives the life of a ghostly Robinson Crusoe. People can see, hear, and jolt one another’s bodies, but they are irremediably blind and deaf to the workings of one another’s minds and inoperative upon them” (Ryle 1949: 13).

Recent developments in cognitive and social neuroscience research have encouraged scientists and their audiences to re-contextualize these features. The new version, still emerging, can be called “Human Nature 2.0” and summed up as follows:

> 1. Mind becomes a visible epiphenomenon of the social brain. It is too early to talk about the relationship between mind and brain in strictly deterministic terms however.
> 2. Humans are rational *agents*, as assumed in Human Nature 1.0. At the same time, minds and brains (the seat of human agency) are *products* of a higher and more stringent kind of rationality, natural selection. A person can be considered rational to the extent that, on a given occasion, her intentions, purposive behavior, and the material results of her goal-directed action are consonant (“sensible”) and proportionate according to the standards of her community. Natural selection is rational in that it is determined a ruthless cost-benefit calculus (reproductive success).
> 3. The “hedonistic calculus” of Human Nature 1.0 is unaffected.
> 4. The most striking difference between the two versions concerns the mechanisms through which minds/brains communicate. In version 1.0, minds know other minds only *indirectly*, through signs and symbols, encoded in language, gestures, and purposive behavior. In version 2.0, there is an additional mechanism: minds are routinely in *direct* contact, via neural resonance, mirroring, and empathy.

THE PREHISTORY OF EMPATHY

The idea that brains and minds might interpenetrate reprises nineteenth century medical discourse and debate on suggestion, hypnosis, and mental contagion. Jean-Martin Charcot claimed that his clinical studies of Mesmerism, hysteria, and psychogenic trauma had led him to believe that the hypnotic state is evidence of a biological diathesis. His claim was contested by Hippolyte Bernheim, who believed that hypnosis is a form of suggestion, and that suggestibility is both universal and normal, notwithstanding the observation that highly suggestible people are more credulous than others. Bernheim defined "suggestion" very broadly, as an "act by which an idea is introduced into the brain and accepted by it." It occurs in two forms: in *hetero-suggestion*, ideas pass from one mind to another; in *auto-suggestion*, ideas emerge spontaneously within the mind, where they become associated with particular sensations, emotions, and images. Auto-suggestion might be a source of distress and even psychosomatic disorders, but it is not intrinsically pathological (Bernheim 1980/1891: 18, 22).

Bernheim's conception of hetero-suggestion was the basis for Gustave Le Bon's influential monograph, *La psychologie des foules* (1895). Le Bon believed that French society was undergoing a massive and unfortunate transformation, that could be traced to the accession of the masses, the *classes populaires*, to political power, The masses want to "utterly destroy society as it now exists, with a view to making it hark back to that primitive communism which was the normal condition of all human groups before the dawn of civilization" (Le Bon 2002/1895: ix-xi; also Nye 1975; van Ginneken 1992: ch. 4). Rationality is a trait of the civilized, autonomous individual. The masses are not individuals in this sense, but rather creatures of a formation called the 'crowd', *la foule*, Once he is part of a crowd, the person "acquires, solely from numerical considerations, a sentiment of invincible power which allows him to yield to instincts which, had he been alone, he would perforce have kept under restraint." His mind and brain become permeable to other minds and brains, and he loses his conscious personality. He now descends the evolutionary ladder. "Isolated, he may be a cultivated individual; in a crowd he is a barbarian…. He possesses the spontaneity, the violence, the ferocity … of primitive beings [and can] be induced to commit acts contrary to his most obvious interests…" (Le Bon 2002/1895: 6 and 8).

Le Bon's accepts Bernheim's thesis that, to varying degrees, everyone is suggestible. However he has no interest in the expression of suggestibility in unexceptional circumstances. And this makes him different from Bernheim. Indeed Boris Sidis, a Harvard psychiatrist and authority on suggestion, criticized Bernheim for defining the trait so broadly as to include most mental activities (Sidis 1898). In practice, Bernheim did very little to challenge the idea of the autonomous, self-contained individual. When he discusses suggestibility, he mentions contagious yawning, the psychosomatic symptoms that can be induced by auto-suggestion, and his efforts to reverse of these symptoms through clinical hetero-suggestion. Human Nature 1.0 remains unaffected.

During the same period, an analogous notion emerged in Germany. Theodor Lipps identified a psychophysical process, *Einfühlung*, superficially similar to Bernheim's notion of 'indirect' suggestion, a spontaneous response to sensory stimuli producing an 'inner imitation'. This is Lipps' description: I observe someone's facial expression of affect. And "there exists within me a tendency to experience in myself the affect that naturally arises from that gesture". When there is no obstacle, the tendency is realized and the subjective meaning of the affect becomes my experience of the affect. *Einfühlung* is 'positive' when it does not conflict with my own character and 'negative' when there is conflict. Even when there is conflict my tendency to experience his affective state remains. Thus a person stares at me in an arrogant way. "I experience within myself the arrogance contained in that look. … My inner being objects; I feel in the arrogant look … a denial of my personality." Within myself, I resist the negative *Einfühlung* and it is this effort contributes (developmentally) to the ontogenesis of the self. It enables subjectivity to separate from the selves that it observes and, so, empathically experiences (Jahoda 1995: 155-159; Pigman 1995: 242-243; Lipps 1903: 193, Pigman's translation).

Lipps' conception of positive *Einfühlung* is similar to the idea of "sympathy" described by earlier writers, notably David Hume and Adam Smith (Penelhum 1993: 134-135). (Lipps had translated Hume for publication in Germany.) In 1909, Edward Titchner introduced Lipps' notion to Anglophone readers as "empathy," but with a significant alteration. Unlike Lipps, he makes an explicit distinction between empathy (the capacity to fully comprehend the situation of the observed individual) and sympathy (the capacity to share the feeling of the observed individual). By the 1930s, Titchner's distinction has

entered psychological discourse and, soon afterward, is absorbed into the everyday language of educated people. The distinction is both analytical and moral. Empathy is a morally neutral state – I comprehend Zande witchcraft beliefs without wishing to promote them. Sympathy readily blends into compassion and perhaps an impulse to improve the situation of the observed individual. Thus the credibility of the self-contained mind is unaffected.

It is a mistake to suppose that ideas about human nature might have evolved differently except for Titchner's interference. Edmund Husserl adopted Lipps' notion of *Einfühlung* for his phenomenology. From the beginning of life, he wrote, human subjectivity comprises *inter-subjectivity*: a relation between self and other in which the other is apprehended by means of a primitive holistic process of "pairing" occurring at the level of the body. But Husserl retains the "primordial ego" as the foundation for this process: he writes about inter-subjectivity without inter-penetration (Moyn 2005: 58-62). Freud mentions Lipps and *Einfühlung* in *Jokes and their Relation to the Unconscious* (1905). By equating *Einfühlung* to the observer's *cognitive* identification with the other's perceptions and intentions, Freud similarly tailors it for a Cartesian ego (Pigman 1995: 244-252).

To summarize: nineteenth-century and early twentieth-century investigations of suggestibility, hypnosis, *Einfühlung*, and empathy did not undermine confidence in Human Nature 1.0 or its representative, the autonomous, self-contained individual. The more serious challenge dates to the 1980s, when it becomes possible, for the first time, to see the mind at work inside the brain.

EMPATHY AND MIRROR NEURONS

Interest in empathy and embodiment has revived as a consequence, in part, of the discovery of the so-called "mirror neurons." The initial mirror neuron research was conducted on rhesus monkeys and utilized an invasive technology permitting scientists to detect and trace the activation of single neurons in the brain's motor cortex. Subsequent research on humans employed non-invasive technologies – most often fMRI – that image the activation of populations of neurons rather than individual cells. In these experiments, the subject observes goal directed behavior being performed by someone else. The sensory input activates a "neural matching system" in the observer's

motor cortex. His activation pattern mirrors the pattern in the performer's brain, and it matches the pattern in his own brain whenever he performs this action. Subjects were asked to passively read action words such as "lick," "pick," and "kick," and fMRI showed mirroring in cortical regions that are activated when tongues, fingers, and feet produce these actions. Similar effects were produced when subjects were asked to imagine themselves or other people performing designated behavior, including expressed emotion. Thus "mirror neurons can be thought of as a sensory-motor gateway for forming *an internal representation of the observed person's state and intents* based on their body language, facial expressions, actions, and so on" (Dinstein 2008: R957, my italics).

Mirror neurons operate in tandem with brain regions and networks responsible for (1) selecting the movements that will be mirrored on a given occasion and (2) inhibiting the performance of the mirrored movements. These two operations are invisible in most laboratory experiments, since they are designed to focus the subject's attention on a single, unambiguous behavior. But life outside the laboratory is more complicated. Multiple actors and actions may simultaneously enter the observer's sensory field. Elements in the field may stimulate imagined events and recall episodic memories each of which can, in turn, become a target for mirroring. Further, many actions remain ambiguous until cognitive processing puts them into context and, only then, makes it possible to infer a goal.

The human neural matching system supports four phenomenological states:

> (1) The observer experiences mirror neuron activation passively in a state called "resonance."
>
> (2) Neural activation engenders a spontaneous and involuntary re-enactment of observed behavior and emotions. This state includes emotional contagion, contagious yawning, and the so-called "chameleon effect."
>
> (3) The observer uncouples his mirrored neural representation and projects it onto its source, i.e. as a cognitive, conative, or emotional state of the individual being observed. The ability to objectify uncoupled representations is called "perspective-taking."
>
> (4) The uncoupled representation is objectified (made explicit) and is accessible to the observer as a resource for "true imitation."

The states are likewise evolutionary and developmental stages. The ability to uncouple mirrored representations (stages 3 and 4) requires the development of structures and networks outside the mirror neuron system. Non-human primates and other mammals get to the second stage, but no further. Normal children are capable of perspective taking and true imitation by the age of four. Perspective-taking is a precondition for "mind-reading." This seems to be a distinctively human capacity that enables us to interpret other people's intentions, predict their behavior, and attempt to manipulate them. (While other mammals lack this ability, there is compelling empirical evidence that some bird species – notably corvids – are adept mind-readers and agents of deception.)

Perspective-taking is the basis for self-conscious empathy. For many writers, mirroring is an intrinsically empathic event, and this view helps to explain the recent explosion of interest in empathy in cognitive and social neuroscience, neuropsychiatry, developmental and evolutionary psychology, anthropology, moral philosophy, evolutionary biology, neuro-economics, neuro-ethics, neuro-aesthetics, and popular science journalism. Here is an excerpt from an article by Daniel Goleman, writing in 2006 in the *New York Times*:

> The fledgling field of social neuroscience is [now] figuring out the brain mechanics [of] the circuitry that underlies the urge to help others in distress. … Mirror neurons operate like a neural WiFi, activating in our own brains the same areas for emotions, movements and intentions as those of the person we are with. This allows us to feel the other person's distress or pain as our own [and we are] moved to help relieve it. Those who feel another's distress most strongly are most likely to help; those less moved can more easily ignore someone else's distress.

Goleman's excerpt reports the consensus view in social neuroscience. It is consistent with Lipps' original notion (*Einfühlung*): the observer can be said to embody the target of his gaze (Carr et al. 2003; Fogassi 2005; Heim and Singer 2008). But the phenomenon goes beyond Lipps' vision. The target's sensory-motor representations have penetrated the observer's brain: the correspondence between brains is identity and not analogy. There is another significant difference with the past. Titchner and later social psychologists made a distinction between empathy and sympathy (compassion). But

Goleman presumes that empathy is not just pro-social**,** it is also morally positive (disposing people to benevolence). This was also the majority view in social neuroscience at the time (2006).

THE SOCIAL BRAIN

The term "social brain" recurs throughout the cognitive and social neuroscience literature. The brain is doubly social: it enables and inclines humans to engage in complex forms of social interaction, and it is the product of our ancestors' five million year adaptation to social life. The two meanings of social are bridged by the brain's capacity for empathy and mind-reading and the biological hardware (notably the mirror neuron system) that serves these functions. The social brain also comprises three evolutionary narratives:

The narrative of the Jacksonian brain

The narrative of other minds

The narrative of the one and the many

The narratives, whose beginnings date back to the seventeenth century, are explorations of the brain's biological and sociological origins, its architecture, its interface with the mind, and the ways in which researchers might penetrate its recesses. They are neither "mere stories" nor the "historical background" to the real business of neuroscience. Because they are an integral to the business, I will want to describe them one by one, with an occasional detour.

1. THE NARRATIVE OF THE JACKSONIAN BRAIN

In the Croonian Lectures on the Evolution and Dissolution of the Nervous System (1884), the neurologist John Hughlings Jackson described the nervous system as comprising a hierarchy of sensory-motor "centers" acquired incrementally as evolutionary adaptations. At the bottom of the hierarchy are the oldest centers – spontaneous, inflexible, reflex-like. The older centers are inhibited and controlled by centers acquired later. When a control center is disabled (by disease, alcohol, etc.), previously inhibited centers are released to perform their evolved functions, and the effect is expressed in symptoms, syndromes, and mental states. These released functions are called "positive" symptoms; a "negative" symptom, such as paralysis, results from the loss of a function. This process, which retraces the nervous system's evolutionary path in reverse order, is called "dissolution." A patient with delirium tremens

who sees non-existent rats and mice is exhibiting a positive symptom consequent to shallow dissolution, leaving several evolutionary layers unaffected. On the other hand, a case of epileptic mania, characterized the explosive discharge of energy and the so-called “dreamy state” that follows grand mal seizures are products of deep dissolution reaching lower evolutionary layers.

Thus the selection of appropriate neuropsychiatric disorders and positive symptoms allows researchers to explore the brain’s evolutionary architecture. Hughlings Jackson’s clinical interest focused on epilepsy and aphasia, and his most extended observations concern these disorders. Following his death in 1911, interest in the Jacksonian brain declined, the exceptions being W.H.R. Rivers in Britain, Paul McLean in the United States (his “triune brain” reiterates the Jacksonian scheme), Henri Ey in France, and arguably Sigmund Freud in *The Interpretation of Dreams*. Interest in the evolutionary meaning of mental disorders reemerged in the 1960s (Price 1967), stimulated by developments in sociobiology and (later) evolutionary psychology. These writers were generally more interested in the architecture of the mind rather than the brain, and their work spanned many conditions, including depression, postpartum depression, antisocial personality disorder, generalized anxiety, schizophrenia, agoraphobia, and animal phobias. In these accounts, each disorder reveals its distinctive evolutionary origin. There is no grand narrative: the mind comes together as a mosaic of evolutionary events and dispositions. The Jacksonian brain is different in this regard. It reemerges (anonymously) in the 1990s, concurrent with the availability of functional neuroimaging technology, the consequent discovery of the human mirror neuron system, and the wide-spread conviction that empathy and mind-reading are core features of human nature and its evolutionary history. To investigate empathy and mind-reading, however, one requires an appropriate assortment of normal and abnormal brains. Three disorders are especially suited to the job: schizophrenia, autism spectrum disorders, and psychopathy.

In one respect, the social brain and Jacksonian brain are quite different. Hughlings Jackson believed that every mental state has a correlative nervous state: the highest link of the purely physical chain of sensory-motor structures. The two states occur in parallel: philosophers of mind call this “property dualism.” He explicitly rejected Descartes’ doctrine of dual substances and likewise “materialists” who claimed that every mental state can be reduced to a discrete neural state. Hughlings Jackson called his position

the doctrine of concomitance. The term is rarely used today, but the problematic – the brain-mind nexus – continues to attract the attention of philosophers, including John Searle, Jerry Fodor, and Daniel Dennett.

Into the 1990s, reductionists lacked an effective technology and research program to bridge mind and brain. This makes the social brain special: it provides a bridge is based on three kinds of empathy, namely motor empathy, emotional empathy, and cognitive empathy. Mirror neurons are a subpopulation of motor neurons that extend to brain regions associated with emotional and cognitive empathy. Thus social brain research has the possibility of delineating a "purely physical chain of sensory-motor structures" extending to the conscious mind, leaping over the doctrine of concomitance. Further evidence is provided by continuing experiments on which participants' brains are imaged while they complete carefully designed cognitive tasks or, alternatively, while they observe emotionally evocative stimuli.

2. THE NARRATIVE OF OTHER MINDS

The size of the human brain is an evolutionary puzzle. Our ancestors split from the great apes six million years ago. During this period, the ancestral human brain quadrupled in volume. The metabolic costs of the human brain are enormous: it constitutes 2% of total body weight and consumes 15% of cardiac output and 20% of body oxygen. These demands are ceaseless and inflexible. A brief shortfall results in neuronal death, resulting in a debilitating and permanent loss in functioning. It can be assumed that the evolutionary growth of the brain reflects an adaptive advantage: the benefits were consistently greater than the metabolic costs. During the initial stage, benefits were caloric and a product of improved adaptation between the organism and the physical environment. Efforts to model the evolution of hominid brains, indicate that increasing costs would eventually exceed environmental benefits. How did the expanding brain pay for itself?

The early history of the hominid brain is about adaptation between organism and physical environment. The subsequent history is about brains adapting to other brains. The process is described as a *cognitive arms race* (Byrne and White 1988; Barton and Dunbar 1997; Dunbar 2003). It began with the emergence of a unique hominid mind-reading capacity: the ability to detect the intentions and predict the behavior of other

members [1]of one's group. The next stage was the emergence of so-called "cheaters," who used mind-reading to manipulate other members. Cheaters would have had an adaptive advantage and therefore multiplied. In time the proportion of cheaters would increase to the point that social life would become unpredictable and regress to the previous, more primitive stage. This did not happen because the brain evolved a "cheater detector" capacity. But this could be only a transient solution, since a new generation of opportunistic individuals would exploit this capacity to cheat a new generation of victims. Once again cheaters thrive, social life grows unpredictable, etc. Devolution is avoided with the emergence of cheater-detector 2.0. And so on, over millions of years, until arriving at the current version of the human brain.

The cognitive arms race is depends on the ability of individuals to detect the intentions of others and predict their behavior – in other words, "mind-reading." The responsible mechanism is the human mirror neuron system, which has its own evolutionary history (Gallese 2000, 2001; Gallese and Goldman 1998; Fogassi et al. 2002 and 2005; Iacoboni et al. 2005; Kohler et al. 2002; Rizzolatti and Arbib 1998; Rizzolatti and Craighero 2004; Tettemonte et al. 2005; cf. Singer 2006; Jacob 2008; Jacob and Jeannerod 2005; see Fadiga et al. 1995 for the discovery of mirror neurons):

> *Stage one*: The observer's mirror neurons *resonate* with the neurons of the agent performing a goal-directed action. A transient "primary representation" of the neural activation pattern is produced in the observer's brain. The brains of non-human primates did not evolve beyond this stage. Emotional contagion is possible, but not emotional empathy.
> *Stage two*: The primary representation can be *uncoupled* from the transient experience and *copied* inside the brain. This is the neural basis for perspective-taking. Cognitive and emotional empathy are now possible.
> *Stage three*: Copies are archived and provide the brain and mind with a library of action patterns. True imitation becomes possible.[2]

The phylogenetic series is replicated in the cognitive development of normal children.

THE PROBLEM OF THE ONE AND THE MANY

Human Nature 1.0 poses an evolutionary puzzle: How did aggregates of autonomous, self-interested individuals – our remote ancestors – coalesce into stable, self-reproducing societies? And once formed, how did the earliest groups evolve into complex social formations?

Thomas Hobbes' thesis was that our ancestors were guided by reason and driven by fear to surrender their private right to use force to a sovereign power that would excercise their strength in the interest of collective peace and defense (Sahlins 2008: 13). Freud's solution in *Totem and Taboo* (1913) is a two-tier hierarchy maintained by the violence and authority of consummately selfish and insatiable patriarch. A parallel solution has been observed among baboons: the hierarchy is stable, the alpha male is similarly violent and insatiable, but the position of individuals within the hierarchy is fluid. John Price, a founding father of evolutionary psychiatry, believes that their situation is very close to the condition of the earliest humans. He sees the legacy of this Paleolithic adaptation in the epidemiology and symptomatology of major depression (Price 1967). Adam Smith offered a third solution. In *The Theory of Moral Sentiments* (1759), he writes that man is doubtlessly selfish, but his self-love is tempered by an imaginative capacity to place himself in the situation of others and by his innate concern for their happiness and misery. This explains the naturalness of pity and compassion. In *The Wealth of Nations* (1776), he responds to the further question of how the earliest groups might have evolve into more complex formations. It is through a human propensity to exchange one thing for another: goods, gifts, and assistance.

The solution given in the evolutionary narrative of the social brain comes close to Adam Smith's account. It emphasizes similar propensities: empathic mind reading and exchange. As you will see, it has problems staying on course.

3. THE NARRATIVE OF THE ONE AND THE MANY

This narrative begins with the riddle of altruism. Population biologists define altruism as behavior in which individuals sacrifice or reduce their own reproductive chances in favor of other members of their group. If this behavior is genetically determined, then altruistic individuals should eventually disappear. This does not occur. The riddle is solved by kin selection theory, which says that altruism is adaptive if the frame of reference is the

survival and reproduction of genes rather than individuals. If so, then altruism is limited to the altruist's relatives, who share some of his genes.

The great leap forward in social evolution is the emergence of reciprocity, a behavior that incorporates non-kin in networks of mutually advantageous exchanges. Like mind-reading, social life evolved dialectically (Bernhard et al. 2006; Boyd et al. 2003; Nowak and Sigmund 2005; Rosas 2008; Simpson and Beckes 2006). Reciprocity creates the possibility of "free-riders." These individuals take but do not reciprocate; they enjoy benefits without costs. The situation recalls the story about deceivers. Non-reciprocators have a reproductive advantage (they get calories without expending energy) and eventually replace reciprocators. Social life regresses. This did not happen because of another evolutionary development: the emergence of punishment in the form retribution or ostracism. Non-reciprocation becomes expensive. Punishment is also expensive for enforcers, who may become targets for retaliation and the disaffection of his own kin and neighbors. Since enforcers jeopardize their own reproductive success, punishment is properly called "altruistic punishment."

Punishment solves a riddle but is also the source of a riddle. Why would a rational individual – someone innately self-interested and capable of calculating cost-benefits – become an enforcer? The enforcer's material benefits are hypothetical. His potential payoff may be in the distant future, and the future costs of his actions are unpredictable. Even if he eventually gets his fair share, he cannot know whether this would have happened without his intervention. Therefore the enforcer's expectation of material rewards can provide only a weak motive for practicing altruistic punishment.

Neuroeconomics – a hybrid of experimental economics and social neuroscience – opened the way to a solution with a landmark experiment, "The neural basis of altruistic punishment," published in the journal *Science* (de Qervain et al. 2004; also Fair and Camerer 2007; Fliessbach et al. 2007; Knoch et al. 2006; Lanzetta and Englis 1989; Singer et al. 2006). The experiment was organized around "the dictator game." One participant is given a sum and told to divide it among other players as he wishes. In subsequent rounds, similar sums are given to the other players. Some players violate cultural standards of fairness and keep an excessive portion for themselves. Participants can punish these so-called "defectors" by withholding payment when the opportunity

arises. However the enforcer must reduce the amount that he pays himself. Thus his behavior is altruistic and pro-social: it contributes to the stability of the network.

Neuroimaging technology (positron emission tomography) was used to observe the enforcers' brains in action. Images showed activation of the caudate nucleus of the dorsal striatum, a "reward center" (pleasure) associated with dopamine excretion. Activation was correlated with the enforcer's *anticipation* of punishing the defector; intensity of activation correlated positively with severity of the punishment. In other words, the enforcer's brain empathically mirrors the imagined (anticipated) distress of the target and, at the same time, delivers pleasure. (The capacity of the brain to mirror imagined distress has been demonstrated in participants who asked to imagine someone else in physical pain (Jackson et al. 2006; see also Singer et al. 2006; Lamm et al. 2007). Parallels with Bernheim's speculations on autosuggestion should be obvious.)

SCHADENFREUDE

The part played by the imagination in the operation of empathic cruelty can be seen directly in a recent study by Takahashi et al. (2009). The study concerns the emotions of envy and *Schadenfreude*, conceived as two sides of one coin. Envy is described as a painful emotion, characterized by feelings of inferiority and resentment, and produced by the individual's awareness of another person's superior quality, achievement, or possessions. *Schadenfreude* is characterized as a pleasurable emotion, produced by awareness that a misfortune has fallen to a person who is envied or otherwise resented.

The Takahashi group recruited nineteen male and female students for their research. Prior to fMRI scans, the participants were asked to read descriptions of three fictive students. (Participants and fictive students were matched for gender.) The first student (A) is the "protagonist": participants are expected to view students B and C from A's perspective. The protagonist is depicted as someone with only average abilities, social endowments, personal achievements, possessions, and prospects. Student B is depicted as someone who is superior and successful in these respects and in the life domains that are important to the protagonist (and participant). Student C is depicted as superior and successful but in domains that are not important to the protagonist.

During fMRI scans, participants silently read scripts pertaining to A, B, and C. The phase one scripts described the successes and advantages enjoyed by B and C. Participants rated the sentences according to how envious the events made them feel (1 = no envy, 6 = extreme envy). Phase two scripts described various misfortunes that spoiled events and prospects for the fictive students. Participants were asked to report the intensity of their pleasure (*Schadenfreude*) regarding each outcome. Thus they provided two responses: subjective appraisals of their emotions, and images of neural activation.

An earlier neuroimaging study (Eisenberger et al. 2003) shoed that physical pain and "social pain" (in the experiment, self-reported distress caused by social exclusion) are associated with the same region of the brain, the anterior cingulate cortex The Takahashi et al. research shows that intense envy (focused on student B) produces a similar activation. On the other side of the coin, intense *Schadenfreude* (likewise focused on student B) is associated with activation of the ventral striatum, described "a central node of reward processing." Thus the *Schadenfreude* effect imaged in this research replicates the events inside the enforcer's brain in the de Quervain et al. study.

EMPATHIC CRUELTY AND HUMAN NATURE

Altruistic punishment persisted throughout the long period following the emergence of social networks based on reciprocity. And it can be assumed that motivation for altruistic punishment was transmitted across generations as a heritable disposition. The rise of state societies, markets, and institutions for regulating exchange reduced the importance of reciprocity and the role of altruistic punishment. But these developments were too recent to affect the disposition to punish, and it can be considered an aspect human nature. In T*he Concise Oxford Dictionary*, "cruelty" is defined as "having pleasure in another's suffering." If so, the disposition can be called *empathic cruelty*.

Recall Daniel Goleman's account of mirror neurons, where he represents empathy as *intrinsically pro-social and morally positive*. This view pervades social neuroscience. "Empathy allows us to understand the intentions others, predict their behaviour, and experience an emotion triggered by their emotion. In short empathy allows us to interact effectively in the social world. It is also the "glue" of the social world, drawing us to help others and stopping us from hurting others" (Lawson et al. 2004: 163; Baron-Cohen et

al. 2005; Wheelright et al. 2006; also Williams et al. 2001 and Iacoboni and Dapretto 2006).

Simon Baron-Cohen, an authority on autism spectrum disorders, writes that human evolution has produced polar types of brains: a female brain with highly developed empathic capacities, and a male brain adapted to manipulating objects and creating systems. Empathy originated as a pro-social adaptation allowing Paleolithic females to detect the wants of pre-verbal children and the moods of the potentially dangerous males with whom they lived. On the other hand, autistic individuals are characteristically poor empathizers. The epidemiology of the disorder is biased towards males: the ratio is 5 to 1, and 10 to 1 with high functioning autistic disorder. We should think of autism as a disorder of the extreme male brain.

According to Baron-Cohen, people respond to suffering in these three ways:

1. The observer's response *mirrors* the sufferer's distress.
2. The observer's response is culturally appropriate but does not mirror the suffering: e.g. the observer responds with sadness to the sufferer's pain.
3. The observer takes pleasure in the sufferer's condition.

Baron-Cohen equates "empathy" – the glue of the social world – with the first two responses. He explicitly excludes the third. He does not consider a fourth possibility, where the observer mirrors the sufferer's distress while taking pleasure in the sufferer's condition. Why? Is "empathic cruelty" a contradiction in terms? De Quervain's research suggests otherwise.

EMPATHIC PSYCHOPATHS

> "Psychopathy can be considered one of the prototypical disorders associated with empathic dysfunction. Reference to empathic dysfunction is part of the diagnosistic criteria of psychopathy. The very ability to inflict serious harm to others repeatedly can be, and is, an indicator of a profound disturbance in an appropriate "empathic" response to the suffering of another." (Blair 2005: 707-8)

Recent research by Jean Decety and his collaborators (n.d.) utilized eight adolescents diagnosed with "aggressive conduct disorder" (CD) and eight matched controls. The

classification “conduct disorder” is limited to young people, generally males. Aggressive CD people have a record of inflicting pain on others. Participants’ brains were scanned with an fMRI apparatus while they watched videos of people experiencing pain resulting from an accident or someone else’s intentional action. Brain images showed that the pain matrix in the CD brains is activated to a significantly greater extent than in the normal brains. They also showed greater activation in the striatum – “part of the system implicated in reward and pleasure.” Regions associated with the *regulation* of emotion were activated to a lesser extent than in the normal brains. And it is assumed that similar activation patterns occur when CD adolescents actually inflict pain on others.

The brain images show that “highly aggressive antisocial youth enjoy seeing their victims in pain and … may not effectively regulate positively reinforced aggressive behavior” – i.e. behavior providing them with “enjoyment” or “excitement.” CD brains and normal brains share an innate capacity for empathic cruelty. The difference between them is that CD brains are *more empathic* than normal brains, but also less capable of regulating the consequent emotion. (This is the favored hypothesis. An alternative hypothesis is that CD youths have a lower threshold for responding to situations of negative affect, including viewing pain in others, and are less able to regulate negative emotion. Distress induces renewed aggression, aggression inflicts more pain, the empathic experience of the pain heightens distress in the CD brain, and so on.)

In the same year, a research team (Fecteau et al. 2008) investigated empathy and psychopathic tendencies in a non-psychiatric population Male college students were asked to watch four videos: a human hand at rest; a Q-tip touching the hand at point X (over the first dorsal interosseus muscle); a needle inserted at point X, and a needle penetrating an apple. During viewing, motor cortex excitation was monitored by transcranial magnetic stimulation (TMS). This technology is able to localize and measure neural responses to pain within the sensorimotor system: muscles at point X mapped onto corresponding regions of the brain. Responses to the static hand video provided a base line. The Q-tip and needle videos elicited reduced motor cortex excitation; the effect was greatest in response to the needle video. The response is characterized as “empathic” (see also Singer and Frith 2005). Participants were also asked to complete a questionnaire, the Psychopathic Personality Inventory (PPI). High scores on “coldheartedness” (callousness, guiltlessness, and lack of sentimentality) correlated with

a greater reduction in cortical excitation.

Fecteau et al. cite research by Avenanti et al. (2005). The Avenanti team followed a similar procedure except that participants were asked to complete a questionnaire measuring “empathic concern” and personal distress. The results are that “massive inhibition of corticospinal excitability affecting upper limb muscles” correlate with high empathy scores: greater reduction equals greater empathy. In other words, while everyone responds empathically to the needle video, the empathic response is more intense in participants with psychopathic tendencies. Thus Fecteau’s team and Decety’s team reach a similar conclusion.
(n.b. This is the explanation for *reduced* neural excitation on these occasions: The response may be part of an evolutionary adaptation that helps the observer’s corticospinal system “implement escape or freezing reactions” (Avenanti et al. 2005: 958).)

In common with Baron-Cohen, the Fecteau team seems reluctant to get to the bottom of the empathic cruelty business. The team visualizes empathy inside the psychopath’s brain, and then asks how one should understand this finding given that “the psychopathic construct … is usually defined by a lack of empathy.” Their solution is to conceive empathy as a two-step process. Step one produces an embodied (mirrored) simulation at a sensory level, facilitates mind-reading, and provides the psychopath with a “substantial advantage for manipulation or harm.” According to DSM-IV, while “deceit and manipulation are general features” of the condition, these individuals “frequently lack empathy” – that is to say, emotional empathy and a benevolent attitude (American Psychiatric Association 1994: 645, 657). According to the Fecteau team, these features, characteristic of true empathy, are produced during step two, when the simulation information needed for mind-reading is made available for an emotional/affective response (pity, sorrow, remorse, outrage, etc.). Thus the exaggerated empathic response that fMRI unexpectedly visualized in the “coldhearted” and participants is explained as the consequence of a defect in step-two processing that “might be maladaptive in psychopaths” (Fecteau et al. 2008: 142).

ON THE GENEALOGY OF MORALITY

Friedrich Nietzsche was less timorous about including empathic cruelty in human nature.

Take a close look at an episode involving cruel punishment, something analogous to the de Quervais experiment, where there is an identifiable perpetrator (enforcer) and victim (non-reciprocator). What are the perpetrator's motives? The obvious answer is that he is gratified by his victim's suffering. But why would the perpetrator's cruelty be gratifying? Well perhaps he believes that his action is pay-back. The victim had previously injured him or deprived him of something to which he is entitled. ("Free-riders" would be perfect candidates for cruelty. But Nietzsche is not thinking in these terms, is unfamiliar with the reciprocity narrative, and cannot imagine "altruistic punishment.") If so, does suffering somehow repay the victim's debt to the perpetrator and reduce his guilt? This "economic" explanation appeals to some moral philosophers, but it makes no sense to Nietzsche. Suffering is not like money. It is not convertible and it cannot be transferred like coins, from one hand to another. The reality, according to Nietzsche, is that the perpetrator is *not* gratified because he sees his victim suffer. He is gratified because the victim's suffering tells him something about himself. It affirms that he, the perpetrator, has *power sufficient to inflict the suffering*. It is the perpetrator's thirst for this knowledge about himself and his world that explains the bond between empathy and cruelty. *It is the perpetrator's ability to empathically experience his victim's suffering, not his ability to exact the suffering from the particular victim, that delivers to him visceral proof of his power* (Nietzsche 1994/1887: 39-43).

In *Crowds and Power* (1962), Elias Canetti deploys Nietzsche's theme to explain Daniel Paul Schreber's paranoid fantasies (*Memoirs of My Nervous Illness*, 1903) and likewise the real enormities perpetrated by Hitler and Stalin. (I will return to Canetti in chapter four.) This theme – connecting empathic cruelty to power and crowds – is also part of Daniel Lord Smail's recent book, *On Deep History and the Brain* (2008) (see also Nell 2006 and Stein 2000). Smail sketches an evolutionary history that begins with our pre-hominid ancestors living in groups resembling chimpanzee and baboon societies. The pre-hominid social order was hierarchical and dominance was maintained by random acts of violence against subordinates. Early hominid society was the next stage. Smail imagines these ancestors living social lives similar to contemporary foragers and hunters. Contra to the situation Freud describes in *Totem and Taboo*, it was relatively egalitarian and presumably less stressful way of life. This lasted until the Neolithic Period. Social hierarchies returned and soon afterward the earliest states emerged.

From the Bronze Age until the early modern times, ruling classes maintained dominance by controlling the bodies and regulating the neurochemistry of the brains of the subservient classes. High levels of stress were maintained among the unfortunates through the exercise of continual and often unpredictable terror and repression. Empathic cruelty provided stress relief, and the masses were periodically treated to sadistic spectacles and horrible priestly rites. But all of this was under the control of the elites. Smail calls these practices “teletropic”: their goal (telos) or function was to preserve the power of the elites.

Conditions changed in Western Europe during the seventeenth and eighteenth-centuries. Alcohol, opiates, sugar, caffeine, tobacco, pornography and sentimental novels became generally available, and now individuals could modulate their own body chemistry. These “autotropic” mechanisms “mimic or alter the effects of dopamine, serotonin, norepinephrine, and other chemical messengers.” It is at this historical juncture, the point at which gin, chocolate, and novels replace empathic cruelty for the relief of stress, that ordinary people are empowered (so to speak) to take control of the caudate nucleus. Fast forward to the twenty-first century: the world of the modern consumer, the apotheosis of the autotropic, the time of *La foule des solitaires.* Thus empathic cruelty endures as part of human nature but is less obtrusive today than it was in pre-modern times, except for hyper-stressful events that habitually intrude into our time in the form of world wars, genocides, ethnic cleansings, economic depressions, etc.

MEMORY AND EMULATION NEURONS

> Mirror neurons offer a tangible integration of perceived and performed action and better still, they do so by means of an experimentally accessible specialized single cell type. The opportunity is hard to resist. (Kinsbourne 2005: 211)

> More than a decade of research on mirror neurons has left us with a crucial problem: is there a mirror neuron system in humans? (Turella et al. 2009: 9)

The standard account of mirror neurons begins with a monkey or human observing a goal-directed action. The observer’s brain mirrors the neuronal activation in the actor’s brain: a representation is registered in the observer’s brain, and the observer brain uses

it to infer the actor's intention (Gallese and Goldman 1998, Iacoboni et al. 2005, Rizzolatti and Craighero 2004). The researcher has a neuroimage that tokens the representation inside the observer's brain. The neural mechanisms that connect the observer's representation to his inference remain to be discovered.

There are other ways to interpret the "mirror effect." Gergely Csibra and colleagues believe that the observer's representation pre-exists the actor's performance (Csibra and Gergeliy 2006; Brass et al. 2007; also Kilner and Frith 2008). It is already there, inside his brain, part of a library of action patterns. The standard view is that mirroring is a response to *any kind* of "goal directed" behavior performed by a conspecific. Csibra cites experiments showing that the observer's brain responds only to *meaningful* goal directed behavior; that is, occasions when a reasonable expectation of reward, such as food in the case of monkeys. When the observed actions involve unfamiliar artifacts, occur in unfamiliar circumstances, or approximate nothing in the observer's repertoire, they are meaningless. They appear to be "goal directed" (purposive) but the goals are obscure, without context. On these occasions, the observer's brain responds by attempting to infer the actor's goal. The brain regions activated and imaged during this process have no mirror properties.

The situation is vastly complicated for human observers and their brains by the fact that most targets of purposive behavior can implicate countless possibilities; that is, they can be elements in multiple, alternative scripts. According to Csibra and colleagues, experiments with adults and preverbal infants provide us with persuasive evidence of an innate human disposition to teleological reasoning, also known as "rational action." An individual observes an actor's meaningful behavior: the meaning is inferred from the context and not from reading the actor's mind. The observer's brain sifts through its repertoire of action patterns, and selects the one that most efficiently serves the inferred goal. The selected pattern (representation) is utilized to anticipate and monitor the actor's behavior. When an actor's behavior is novel but successful, the observer brain modifies the
action pattern and adds it to the repertoire of patterns. In other words, the "mirror neurons" that are being imaged in similar experiments are more correctly called "emulation neurons" ("to equal or exceed"). Emulation neurons and goal-oriented rationality would play a pro-social role in human evolutionary history. One can also

imagine that they would be the basis for an alternative, less exciting, version of the cognitive arms race. In other words, they would have brought us to where we are today. What one cannot reasonably imagine is that emulation neurons and goal-oriented rationality would provide a platform for the emergence of intersubjectivity – brains penetrating other brains – or empathic cruelty, or passage to Human Nature 2.0.

Ilan Dinstein and his collaborators have questioned the existence, or at least the pervasiveness, of human mirror neuron system from another perspective. Mirror neurons are presumed to respond during the execution of a specific movement and also when the individual observes this movement. "[C]ross-modal adaptation is the critical signature of mirror neurons since visual adaptation may also be generated by movement-selective visual neurons … and motor adaptation may be generated by movement-selective motor neurons." The fMRI technique measures the *average* neural response across a large neural population located within each voxel (a volumetric sector within a neuroimage), and it is difficult to separate the relative contribution of the different neural subpopulations within a given voxel (Dinstein 2008: R957). There is an alternative way to explain the attributed "mirror" effect within a voxel that is equally consistent with the averaged response. "Such a response could be generated by separate visual and motor neural populations co-existing in the same voxels. Even more importantly, such responses could be generated by neural populations that are not movement selective at all," but are associated with a variety of motor behaviors (Dinstein 2008: R958). A recent experiment, designed specifically to detect a mirror effect from among averaged effects, found no evidence of the presumed cross-modal adaptation (i.e. selective visuomotor neurons) anywhere in the brain (Dinstein et al. 2008).

Csibra's thesis and Dinstein's conclusion are controversial and, for the moment at least, have been swept to the margins of attention by the momentum of the human mirror neuron discourse.

Homunculi

I began the chapter with the claim that Human Nature 1.0 and 2.0 are configured to solve two problems: the problem of other minds and the problem of the one and the many. Human Nature 2.0 solves the problems via an evolutionary pathway that can be summarized as follows:

1. The human brain and its precursor, the hominid brain, automatically mirror the goal-directed behaviour and emotions observed in others.
2. Neuronal mirroring creates representations that enable human observers to know the intentions of others (mind-reading).
3. The biological evolution of human and hominid brains is dialectically linked to the evolution of human and hominid society.
4. Empathy is the subjective experience of mirroring and it likewise has an evolutionary history. Empathy is prosocial in that, like mind-reading, it contributed to social evolution and the stabilization of social groups.
5. The credibility of this account is based on evidence from experiments, neuropathy and mental disorders, and population modeling. The most persuasive evidence is based on neuro-images. But these images are open to conflicting interpretations and are the source of three controversies.

There are neuroscientists, such as Csibra and Dinstein, who question the *existence* of either human mirror neurons or a human mirror neuron system.

Other critics question the *completeness* of the mirror-neuron theory. The experimental situation is designed to focus the participant's attention on a single target. In everyday life, an individual's perceptual field often includes multiple, competing targets. In addition, neuroscientists have repeatedly demonstrated that the brain mirrors targets positioned in memory, imagination, and language. Perceptions are not spontaneously meaningful. They must be interpreted: embedded in webs of associations (via memory). The activation of associations is the source of targets for automatic mirroring. This is also an everyday situation. The experimental situation is designed to de-contextualize its targets and creates the illusion that perceptions and representations are routinely produced without interpretation. How then does the brain mirror and successfully represent everyday situations?

Finally, the human mirror neuron system has been criticized as *incoherent*. The theory behind the system is based on the idea that mirrored representations in the brain are automatically translated into mental representations accessible to consciousness. This is the basis of the brain's mind-reading ability and the evolutionary process responsible for the social brain, and it is a refutation of the doctrine of concomitance advocated by

Hughlings Jackson (1887). To understand the critics' perspective, we should return to Gilbert Ryle's claim that, in the cultural imaginary, the human mind is perceived as a self-contained place in which "each of us lives the life of a ghostly Robinson Crusoe." Daniel Dennett described the place as a "Cartesian theater" within which a ghostly inner-self watches representations of situations encountered by the bodily outer-self. Inner-self and outer-self are near-duplicates.

The cultural ideal is an inner-self that distinguishes between its own inner nature and the artifice of its self-representations – the face that it turns to the social world. In other words, the ideal inner-self is without self-deception, false consciousness, or pathological distortion. Self-awareness is the tip of the pyramid of consciousness, and beneath it essential cognitive operations are continually running.

According to Jerry Fodor (2000, 2006), the operations would include a way to resolve the "frame problem". The human mind (and its brain) must manage a colossal knowledge-base ("cognitive commitments"), stored in episodic memory, semantic memory, and implicit memory. How does it determine what is relevant on a given occasion for identifying an ambiguous object or action, or deciding on a new belief or plan of action? "There is an infinite corpus of prior cognitive commitments that might prove germane, but one can actually visit only some relatively small, final subset of them in the 'real time' during which problems get solved." When cognitive scientists suggest that the problem can be managed by resorting to heuristic cognitive strategies, they simply defer the problem, since everything now depends on selecting one from many available heuristics." Further, "how one individuates situations depends on what one takes as relevant to deciding when situations are the same kind," i.e. past and present. It is likewise unclear how the mind performs its characteristic mode of inference, a rule-less process called "abduction" and "argument to the best available explanation" (Fodor 2006: 90, 93)?

The experimental situation in brain-imaging research has been designed to circumvent the frame problem. How is the frame problem managed in everyday life? According to Fodor:

> [Cognitive science] doesn't actually know how the mind works. Nor do I. Nor does anybody else. And I suspect such is the state of he art that, if God were to tell us how it works, none of us would understand Him" (Fodor 2002: 94).

Human Nature 1.0 permits two possibilities, both. There is the Cartesian response: the frame problem is managed by a *res cogitans* (homunculus) inaccessible to self-awareness. There is Fodor's homunculus-free response: operations are performed by a "central processor." The Cartesian solution is obviously flawed. In order for a homunculus (*res cogitans*) to know how to manage the frame problem, it would need its own homunculus. Thus an endless regression from homunculus down to homunculus. Fodor's idea of a central processor is equivalent to an admission of ignorance: we know the operations that the mind must be performing but no one knows how they are performed.

In the world of Human Nature 2.0, the transformation of neuronal representations into cognitive representations is performed computationally by inter-connected, self-adjusting physical systems. The frame problem is nowhere in sight. There is a different problem: a "representation" is simultaneously a representation *of* something and a representation *for* something, in the sense that a map (representation) presumes the presence of a map-reader: "nothing is intrinsically a representation of anything; something is a representation only for or to someone; any representation or system of representation requires at least one user of the system who external to the system" (Fodor 1978: 101). Who is the map-reader in the social brain?

Daniel Dennett proposes a homunculus solution fit for Human Nature 2.0 and the human mirror neuron system. The out-dated vision is an endless string of intelligent homunculi anchored to first person consciousness: that is, each homunculus relies on its own homuncular *res cogitans*. Dennett is proposing something radically different: a system composed of *hordes of nesting homunculi*.

> The looming infinite regress can be stopped … not by abandoning the basic the basic idea but by softening it. As long as your *homunculi* are more stupid and ignorant than the intelligent agent that they compose, the nesting of homunculi

> within homunculi can be infinite, [but] bottoming out eventually, with agents so unimpressive [i.e. stupid] that they can be replaced by machines (Dennett 2001: 225; described elsewhere as “robots”).

Inside this brain, there is no Self or Subject in the old sense, that is no Subject as witness, higher executive, or map reader. They are all gone – “broken down into subcomponents that are themselves *clearly* just unconscious underlaborers which themselves work … without supervision.” So imagine the process in which neural representations are transformed into cognitive representations as “the workings of a vacant automated factory – not a Subject in sight…” (Dennett 2001: 228-229).

METAPHORS AND MECHANISMS

Very nice and very metaphorical. The homunculi, the underlaborers, and the vacant factory stand in for operations presumably performed by neural mechanisms. The attractiveness of Dennett’s account is, as he implies, vindicated by the neuroscience research that uncovered these mechanisms. Thus Fodor is left behind, still in the twentieth century. But is this truly the case – the part about non-metaphorical physical mechanisms? In this section, I describe the most comprehensive effort so far to identify mechanisms responsible for translating mirrored representations. The putative mechanisms turn out to be the *functions* that a metaphorical system must perform in order for the organism to read minds etc. And the terminology derives not from biology, but rather cognitive science, artificial intelligence, and robotics.

Mirror neurons operate through multiple sensory modalities. Simple activation patterns, such as the act of reaching plus grasping, constitute “pre-wired intentional chains.” The developing brain connects the simple representations to form more complex representations, and these behavioral modules are strung together into programs (Hurley 2005: 185). The elements composing these complex chains are “logically related.” Access to this logic allows observers to intuit actors’ intentions (Iacoboni et al. 2005). Gallese (2003) believes that this logic – in which causality is internally represented – is enabled by the brain’s *forward model architecture*, a figurative structure that researchers locate in the lateral cortex of the cerebellum and projecting into the parietal lobes. Forward model architecture controls motor output and regulates bodily movement in all vertebrate species:

> When I am going to stretch my arm to grasp a handle in front of me, the resulting postural perturbation that would follow, causing my body to bend [forward], is canceled by a forward signal sent to the posterior muscles of my leg, which stabilize my standing posture. The muscles … contract before my arm is set in motion. The contraction … anticipates, *predicts* the outcome of the programmed action of the arm, [the] perturbation, [thus] preventing it…. Neither overt knowledge nor conscious inference is involved… (Gallese 2001: 38).

The forward architecture is the foundation of the mind-reading system. It is part of a sensorimotor loop composing a circuit. Within the circuit, an *inverse model* relays internal sensory feedback permitting the system to adjust its predictions and performance. Internal representations permit the system to compare the predictions (forward model) with the internal sensory feedback (inverse model). After many cycles, the system stabilizes and it is unnecessary to continue to compare and adjust output. The circuit speeds up and the organism's efficiency is enhanced. If a significant mismatch develops between real-time feedback and the system's predicted ("simulated") feedback, the model readjusts the motor command and modifies its predications (simulation). The system again stabilizes.

A fly in the ointment? The creation of manageable internal representations in the laboratory is carefully controlled, but outside the laboratory stimuli arrive from multiple sources in the ambient environment and from memory and imagination. The task is to maximize the sound-to-noise ratio; the solution is to insert "filters" into the circuit.

In humans, the forward model architecture contributes to the emergence of a self. This is how it happens: Mirror neurons are sensory-motor neurons, thus sensory input (e.g. visual perception) induces the motor activation pattern that is observed in the target. The internal representation produced must be put off-line. Otherwise, observers would automatically imitate observed behavior. This operation has important phylogenetic and ontogenetic consequences, in that the system's ability to inhibit the behavior presumes the system's capacity to identify whether a movement (visual feedback) is the target's or the observer's. The subjective distinction that humans make between "self" and "other"

emerges from this operation (Hurley 2005: 187-8). Mind-reading and empathy become possible.

According to Gallese, the internal representations as neither image-like nor are they propositional in the conventional sense (based on analogy with sentences). He describes them as “non-propositional concepts” that encapsulate formal properties of objects and relations between objects. Non-propositional concepts can be manipulated to make inferences and judge probability. Rick Grush (2004) proposes something similar, “amodal spatial imagery,” neither picture-like nor obviously propositional. This format represents objects according to their location and motion within a dynamic framework, encapsulating the brain’s forward architecture. The system anticipates and responds to internal forces and resistances and to feedback from a dynamic world filled with forces and resistances. (The concept is derived from the “dynamic model” outlined in Schwartz 1999 and elaborated in Kosslyn 2005.) Thus one can say that the system has an inner logic (but distinct from “logic” and “logical entailment” as conventionally understood).

According to Gallese, the brain’s forward architecture is the source of our ability to intuit intentions from internal representations. Peter Gärdenfors, a Swedish cognitive scientist, likewise traces the human capacity for causal reasoning to the forward architecture (2004: 403). And he cites research by Povinelli (2000) and Tomasello (1999) showing that apes are bad at reasoning about physical causes and cannot understand intentionality in others. Intermediate and hidden forces are also unknown to them. “On the other hand, even small children show strong signs of interpreting the world with the aid of hidden forces and other causal variables.” According to Gärdenfors, one can assume that human brains possess a more evolved forward architecture than non-human brains.

Recent research into delusions of alien control has produced similar findings (Blakemore et al. 2003). These delusions are characteristic of schizophrenic people who misattribute self-generated actions to an external source. The researchers induced similar misattributions in normal people via hypnotic suggestion. Using PET scan technology, they compared the neural correlates of actions that participants correctly attributed to themselves and movements they misattributed to an external source. Misattributed

movements were associated with higher activation of the cerebellum and parietal cortex. The cerebellum is linked to predicting sensory consequences of movement and it is the presumed location of forward model architecture. Normally, the architecture distinguishes between internal sensory feedback and sensory feedback from exogenous sources. When the function fails, internally generated movements are attributed to external causes. Thus the failed function – induced in the research subjects, occurring spontaneously in schizophrenic individuals – explains the delusions of external influence.

The evidence seems overwhelming and it is easy to forget that the indicated "mechanisms" are really metaphors, functions, and logics, and locating an function in some part of the brain is not equivalent to describing a mechanism. There is nothing intrinsically unscientific about employing metaphoric entities for explaining mental and behavioral phenomena. (See Machamer et al. 2000: 5-8 and 13-15 concerning "ontic adequacy."). It is hard to imagine how else one might proceed in these circumstances, explaining how "internal representations" can be translated into inferences. The point at hand is that findings grounded on Human Nature 1.0 (a central processor inclined to analogical reasoning) and Human Nature 2.0 (a human mirror neuron system) are no different in this regard, notwithstanding the headlines in *the Science Times*.

THE FOLLOWING CHAPTERS

> Professional athletes and coaches … have long exploited the brain's mirror properties perhaps without knowing their biological basis, Dr. Iacoboni said. Observation directly improves muscle performance via mirror neurons. Similarly, millions of fans who watch their favorite sports on television are hooked by mirror neuron activation. … In yet another realm, mirror neurons are powerfully activated by pornography, several scientists said. For example, when a man watches another man have sexual intercourse with a woman, the observer's mirror neurons spring into action. The vicarious thrill of watching sex, it turns out, is not so vicarious after all.
>
> Sandra Blakeslee , *New York Times*, January 2006

In the following chapters, I have something to say about the popular appeal of Human Nature 2.0 and its mirror neuron system. Is an explanation really necessary? How could it be otherwise, given the immense cultural authority now attached to neuroscience? How can one resist the researchers' aesthetic, tied to an unprecedented ability to show the mind working inside the brain, in real-time and color? One is simply overwhelmed by the wealth and diversity of evidence that streams from cognitive neuroscience, social neuroscience, neuro-economics, neuro-law, neuro-ethics, neuro- aesthetics, neuro-marketing, and so on. More than that, this evidence is *intuitively* right. It's as if the fMRI images are just confirming what we already know. And it is this intuition that interests me in the next pages. You will understand that I am not criticizing these intuitions. I just want to get to the bottom of things, to what is taken for granted. And this is my thesis:

1. Mirroring does not explain explains empathy, rather empathy explains mirroring. Empathy is the glue that holds society together, Baron-Cohen writes. What is more certain is that empathy is holds the social brain together.

2. The operations of the empathic brain are anticipated in a range of phenomena in which people are said to compulsively replicate what they see, imagine or remember. This includes mental contagion, mass hysteria, conversion hysteria, autosuggestion, hypnotic suggestion, transference neurosis, projective identification, the chameleon effect, and the repetition compulsion (traumatic neurosis). A second list would include types of conscious simulation and dissimulation – counterfeits, frauds, mimics – in which the performer not only imitates a target but also internalizes it, to some extent.

3. Empathy has a history – not just the concept "empathy," but likewise the empathic sensibility that we take for granted. This empathic sensibility evolved during the second half of the twentieth century – a trajectory recalling the "civilizing process" (two centuries earlier) described by Norbert Elias (1969).

There are many ways to write a history of empathic sensibility. My approach is through the lens provided by pathology: more precisely, a sequence of psychopathologies traced to trauma and facsimiles of trauma beginning in the 1880s. In the intervening years, a wide variety of traumatic images and motifs have seeped into the cultural imaginary. The

process began with influential clinical narratives and monographs (mainly but not exclusively psychoanalytical), followed by a stream of popular novels, films, theatre pieces, memoirs, and journalistic accounts, and currently well-served by confessional television, countless discourses by "trauma theory" experts in comparative literature, cultural studies, film studies etc., and an avalanche of empirical studies by epidemiologists, clinical and experimental psychologists, social scientists and historians. The vision of trauma that has emerged is intrinsically empathic and mimetic, evidenced in clinical conditions and popular tropes associated with contagion (the "vicarious PTSD" that afflicts psychotherapists, the "second-generation PTSD" that afflicts children of trauma survivors); repetition of the past ("flashbacks," symptomatic "re-experiencing"); self-suggestion ("factitious traumatic memory"), dissimulation ("fictitious memory"), and a resonant kind of empathic cruelty in the style of Lady Macbeth ("self-traumatized perpetrators").

REFERENCES

Adolphs, R. 2003. Cognitive neuroscience of human social behaviour. *Nature Reviews: Neuroscience* 4: 165–178.

American Psychiatric Association.1994. *Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association – Fourth Edition*. Washington, D.C.: American Psychiatric Association.

Axelrod, R. and W.D. Hamilton. 1981 The evolution of cooperation. *Science* 211: 1390-1396.

Baron-Cohen, Simon, Rebecca C. Knickmeyer, and Matthew K. Belmonte. 2005. Sex differences in the brain: implications for explaining autism. *Science* 310: 819-823.

Barton, R.A. and R.I.M. Dunbar. 1997. Evolution of the social brain. In *Machiavellian Intelligence II*, ed. A Whiten and R.W. Byrne, pp. 240-263. Cambridge: Cambridge Univ. Press.

Bernhard, H., U. Fischbacker, and E. Fehr. 2006. Parochial altruism in humans. *Nature* 442: 912-915.

Bernheim, H. 1891. *Hypnotisme, suggestion, psychothérapie: Etudes nouvelles*. Paris: O. Doin,

Blair, R.J.R. 2005. Responsing to the emotions of others : Dissociaing forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition* 14 : 698-718.

Blakemore, S.-J., P. Fonlupt, M. Pachot-Clouard, C. Darmon, P. Boyer, A.N. Meltzoff, C. Segebarth, and J. Decety. How the brain perceives causality: an event-related fMRI study. *NeuroReport* 12: 3741-3746.

Blakemore, S.-J., D.A. Oakley, and C.D. Frith. 2003. Delusions of alien control in the normal brain. *Neuropsychologia* 41: 1058-1067.

Blakeslee, S. 2006. Cells that read minds. *New York Times*, 10 January 2006.

Boyd, R., H. Gintis, S. Bowles, and P.J. Richerson. 2003. The evolution of altruistic punishment. *Proceedings of the National Academy of Science* 100: 3531-3535.

Brass, M., R.M. Schmitt, S. Spengler, and G. Gergely. 2007. Investigating action understanding: inferential processes versus action simulation. *Current Biology* 17: 2117–2121.

Brüne, M., H. Ribbert and W. Schiefenhövel, eds. 2003. *The Social Brain: Evolution and Pathology*. London: Wiley.

Byrne, R.W. and A. Whiten, eds. 1988. *Machiavellian Intelligence*. Oxford: Oxford Univ. Press.

Camerer, C.K. and E. Fehr. 2006. When does "economic man: dominate social behavior? *Science* 311: 47-52.

Canetti, E. 1962. *Crowds and Power*. London: Gollanz.

Canguilhem, G. 1978 (orig. 1966). *The Normal and the Pathological*. Cambridge, MA: MIT Press.

Carr, L., M. Iacoboni, M.-C. Dubeau, J.C. Mazziotta, and G.L. Lenzi. 2003. Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Science* 100: 5497-5502.

Csibra, G. and G. Gergely. 2007. 'Obsessed with goals': functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica* 124: 60–78.

Daston, L. 1988. Classical Probability in the Enlightenment. Princeton: Princeton University Press.

n.d. Decety, J., K.J. Michalska, Y. Akitsuki, B.B. Lahey. 2008. Atypical empathic responses in adolescents with aggressive conduct disorder: a functional MRI investigation. *Biological Psychology* in press.

Dennett, D. 2001, Are we explaining consciousness yet? *Cognition* 79: 221-237.

de Quervain, D. J.-F., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, and E. Fehr. 2004. The neural basis of altruistic punishment**.** *Science* 305: 1254-1258.

de Vignemont, F. and T. Singer. 2006. The empathic brain: how, when and why? *Trends in Cognitive Sciences* 10: 435-441.

Dunbar, R.I.M. 2003. The social brain: mind, language, and society in evolutionary perspective. *Annual Review of Anthropology* 32: 163-181.

Eisenberger, N.I., M. D. Lieberman, K. D. Williams. 2003. Does rejection hurt? An fMRI study of social exclusion. *Science* 302: 290-292.

Elias, N. 1969 (orig. 1939). *The Civilizing Process*. Oxford: Blackwell.

Fadiga, L., L. G, Pavesi, and G. Rizzolatti. 1996. Motor facilitation during observation: a magnetic simulation study. *Journal of Neurophysiology* 73: 2608-2611.

Fehr, E. and C. F. Camerer. 2007. Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Science* 11: 419-427.

Fliessbach, K., B. Weber, P. Trautner, T. Dohmen, U. Sunde, C. E. Elger, A. Falk. 2007. Social comparison affects reward-related brain activity in the human ventral striatum. *Science* 318: 1305-1308.

Fodor, J. 2006. How the mind works: what we still don't know. *Daedalus* 35 (3): 86-94.

Fodor, J. 2000. *The Mind Doesn't Work that Way: the Scope and Limits of Computational Psychology*. Cambridge, MA: MIT Press.

Fodor, J. 1975. *The Language of Thought*. New York: Crowell.

Fogassi, L., P.F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, and G. Rizzolatti. 2005. Parietal lobe: from action organization to intention understanding. *Science* 308: 662–667.

Fogassi, L., V. Gallese, and G. Rizzolatti. 2002. Hearing sounds, understanding actions: action representation in mirror neurons. *Science* 297: 846-848.

Gallese, V. 2006. Intentional attunement: A neurophysiological perspective on social cognition and its disruption in autism. *Brain Research* 1079: 15-24.

Gallese, V. 2003. A neuroscientific grasp of concepts: from control to representation. *Philosophical Transactions of the Royal Society of London B* 358: 1231-1240.

Gallese, V. 2001. The "shared manifold" hypothesis: from mirror neurons to empathy. *Journal of Consciousness Studies* 8: 33-50.

Gallese, V. and A. Goldman. 1998. Mirror neurons and the simulation theory of mind reading. *Trends in Cognitive Science* 2: 493–501.

Greene, J. and J. Haidt. 2002. How (and where) does moral judgment work? *Trends in Cognitive Sciences* 6: 517-523.

Grush, R. 2004. The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences* 27: 377-398, 425-442.

Hein, G. and T. Singer. 2008. I feel how you feel but not always: the empathic brain and its modulation. *Current Opinion in Neurobiology* 18: 1–6.

Horton, R. 1967. African traditional thought and Western science. *Africa* 37: 50-71, 155-187.

Hurley, S. The shared circuits hypothesis: a unified functional architecture for control, imitation, and simulation. In *Perspectives on Imitation: From mirror neurons to memes,* ed. S. Hurley & N. Chater, pp. 177-193. Cambridge MA: MIT Press.

Iacoboni, M. and M. Dapretto. 2006. The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience* 7: 942-951.

Iacoboni, M., I. Molnar-Szakacs, V. Gallese, G. Buccino, J.C. Mazziotta, and G. Rizzolatti 2005. Grasping the intentions of others with one's own mirror neuron system. *PLoS* [*Public Library of Science*] *Biology* 3: e79.

Jacob, P. 2008. What do mirror neurons contribute to human social cognition? *Mind and Language* 23: 190-223.

Jacob, P. and M. Jeannerod. 2005. The motor theory of social cognition: a critique. *Trends in Cognitive Sciences* 9: 21-25.

Jahoda, G. 2005. Theodor Lipps and the shift from "sympathy" to "empathy." *Journal of the History of the Behavioral Sciences* 42: 151-163.

Johnson, M., *R. G*riffin, *G. C*sibra, H. Halit, T. Farroni, M. de Haan, L.A. Tucker, S. Baron–Cohen, and J. Richards. 2005. The emergence of the social brain network: evidence from typical and atypical development. *Development and Psychopathology* 17: 599–619.

Kass, L. 2003. Session on Human Nature and Its Future. President's Council on Bioethics. http://www.bioethics.gov/transcripts/march03/session3.html, accessed on 16 April 2009.

Kilner, J.M. and C.D. Frith. Action observation: inferring intentions without mirror neurons. *Current Biology* 18: R32-R33. 2008

Kinsbourne, M. 2005. Imitation as entrainment: brain mechanisms and social consequences. In *Perspectives on Imitation: From mirror neurons to memes,* ed. S. Hurley & N. Chater, pp. 163-172. Cambridge MA: MIT Press.

Knoch, D., A. Pascual-Leone, K. Meyer, V. Treyer, and E. Fehr. 2006. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* 314: 829-832.

Kohler, E, C. Keysers, M.A. Umlitá, L. Fogassi, V. Gallese, and G. Rizzolatti. 2002. Hearing sounds, understanding actions: action representation in mirror neurons. *Science* 297: 846-848.

Kosslyn, S.M. 2005. Mental images and the brain. *Cognitive Neuropsychology* 22: 333-347.

Lanzetta, J.T. and B.G. Englis. 1989. Expectations of cooperation and competition and their effects on observers' vicarious emotional responses. *Interpersonal Relations and Group Processes* 56: 543-554.

Lawson, J., S. Baron-Cohen, and S. Wheelwright. 2004. Empathising and systemising in adults with and without Asperger Syndrome. *Journal of Autism and Developmental Disorders* 34: 301-310.

Le Bon, G. 2002. *The Crowd: A Study of the Popular Mind*. Mineola NY: Dover.1895, translation of Le Bon, G. 1895. *La psychologie des foules*. Paris: Alcan.

Lieberman, M.D. and N.I. Eisenberger. 2009. Pains and pleasures of social life. *Science* 323: 890-891.

Lloyd, G.E.R. 1990. *Demystifying Mentalities*. Cambridge: Cambridge University Press.

Lukes, S. 1973. *Individualism*. Oxford: Blackwell.

Machamer, P., L. Darden, and C.F. Craver, 2000, Thinking about mechanisms. *Philosophy of Science* 67: 1-25.

Moyn, D. 2005. *Origins of the Other: Emmanuel Levinas between Revelation and Ethics*. Ithaca: Cornell University Press.

Nell, V. 2006. Cruelty's rewards: the gratifications of perpetrators and spectators. *Behavioral and Brain Sciences* 29: 211-224; 246-257.

Nietzsche, F. 1994 (orig. 1887). *On the Genealogy of Morality*. Cambridge: Cambridge Univ. Press.

Nowak, M.A. and K. Sigmund. 2005. Evolution of indirect reciprocity. *Nature Reviews Neuroscience* 437:1291-1298.

Nye, R.A. 1975. *The Origins of Crowd Psychology: Gustave Le Bon and the Crisis of Mass Democracy in the Third Reoublic*. London: Sage.

Ochsner, K.N., Z. Jamil, J. Hamelin, D.H. Ludlow, K. Knierirm, T. Ramachandran, G. H. Glover, and S.G. Mackey. 2008. Your pain or mine? Common and distinct neural systems supporting the perception of pain in self and other. *Social Cognitive and Affective Neuroscience* 2: 144-160.

Penelhum, T. 1993. Hume's moral psychology. In *The Cambridge Companion to Hume*, ed. D.F. Norton, pp. 117-147. Cambridge: Cambridge Univ. Press.

Pigman, G.W, 1995. Freud and the history of empathy. *International Journal of Psycho-Analysis* 76: 237-256.

Price. J.P. 1967. The dominance hierarchy and the evolution of mental illness. *Lancet* ii: 243-246.

Rizzolatti, G. and M. A. Arbib. 1998. Language within our grasp. *Trends in Neuroscience* 21: 188-194.

Rizzolatti, G. and L. Craighero. 2004. The mirror-neuron system. *Annual Review of Neuroscience* 27: 169–192.

Rosas, A. 2008. The return of reciprocity: a psychological approach to the evolution of cooperation. *Biology and Philosophy* 23: 555–566.

Sahlins, M. 2008. The Western Illusion of Human Nature. Chicago: Prickly Pear Press.

Schwartz, D.L. 1999. Physical imagery: kinematic versus dynamic models. *Cognitive Psychology* 38: 433-464.

Sidis, B. 1898. *Psychology of Suggestion: Research into the Subconscious Nature of Man and Society*. New York: Appleton.

Simpson, J.A. and L. Beckes. 2006. Reflections on the nature (and nurture) of cultures. *Biology and Philosophy* 23:257–268.

Singer, T. 2006. The neuronal basis and ontogeny of empathy and mind reading: review of literature and implications for future research. *Neuroscience and Biobehavioral Reviews* 30: 855-863.

Singer, T., B. Seymour, J.P. O'Dougherty, K.E. Stephan, D.J. Dolan, and C.D. Frith. 2006. Empathic neural responses are modulated by the perceived fairness of others. *Nature* 439: 466-469.

Smail, D. L. 2008. *On Deep History and the Brain*. Berkeley: University of California Press.

Smith, C.U.M. 1982. Evolution and the problem of mind: Part I. Herbert Spencer. *Journal of the History of Biology* 15:55-88.

Stein, D.J. 2000. The neurobiology of evil: psychiatric perspectives on perpetrators. *Ethnicity and Health* 5: 303-315.

Takahashi, H., M. Kato, M. Matsuura, D. Mobbs, T. Suhara, and Y. Okubo. 2009. When your gain is my pain and your pain is my gain: neural correlates of envy and Schadenfreude. *Science* 323: 937-939. Supporting Online Material, at www.sciencemag.org/cgi/content/full/323/5916/937/DC1

Taylor, C. 1989. *Sources of the Self: The Making of Modern Identity*. Cambridge MA: Harvard Univ. Press.

Tettamanti, M., G. Buccino, M.C. Succaman, V. Gallese, M. Danna, P. Scifo, F. Fazio, G. Rizzolatti, S.E. Cappa, and D. Perani. 2005. Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience* 17: 273-2781.

Trivers, R.L. 1971. The evolution of reciprocal altruism. *Quarterly Review of Biology* 46: 35-57.

Trout, J.D. 2008. Seduction without cause: uncovering explanatory neurophilia. *Trends in Cognitive Sciences* [article in press].

Turella, L., A.C. Pierno, F. Tubaldi, and U. Castiello. 2009. Mirror neurons in humans: conisisting or confounding evidence? *Brain and Language* 108: 10-21.

Van Ginneken, J. 1992. Crowds, Psychology, and Politics 1871 – 1899. Cambridge: Cambridge Univ. Press.

Visalbergh, E. and D. Fragaszy. 2002. "Do monkeys ape?" – ten years after. In *Imitation in Animals and Artifacts*, eds. K. Dautenhahn and C.L. Nehaniv, pp. 471-499 Cambridge, MA: MIT Press..

Wheelwright, S., S. Baron-Cohen, N. Goldenfeld, J. Delaney, D. Fine, R. Smith, and A. Wakabayashi. 2006. Predicting Autism Spectrum Quotient (AQ) from the Systemizing Quotient-Revised (SQ-R) and Empathy Quotient (EQ). *Brain Research* 1079: 47-56.

Williams, J.H.G., A. Whitten, T. Suddendorf, and D.I. Perrett. 2001. Imitation, mirror neurons and autism. *Neuroscience and Biobehavioral Review* 25: 287-295.

Winch, P. 1964. Understanding a primitive society. *American Philosophical Quarterly* 1: 307-324.

---

[2] Some researchers cite evidence that monkeys are capable of imitation. See Visalberghi and Fragaszy 2002 for a review of this research and their reason for rejecting this evidence: true imitation is a form of social learning, monkeys are capble of social learning, but not imitation. According to the authors, evidence for (limited) imitation by apes is more convincing.