JOHANNES HEIDEMA & WILLEM LABUSCHAGNE

Emancipating Agents: Need Schrödinger's Cat be let into the Chinese Room?

INTRODUCTION

In reflecting on the human condition, one has to face the Anthropocentric Question: Are humans really special? Most of us would like to believe the affirmative, and look to science for support.

Carl Linnaeus classified humans with monkeys and apes in 1758, and humans have felt insulted ever since. For two centuries we managed to console ourselves with the thought that human language, intellectual capacity, and culture made us more special than Linnaeus's classification acknowledged. In recent years, however, the rise of cognitive science as a laboratory for testing philosophical ideas has furnished new perspectives on the human condition by situating us within a spectrum of various sorts of agents, and the insights to which this has led do not confirm what we wish were true. Recent work on bonobos[1] and chimpanzees[2] has largely undermined the consolatory belief in an unbridgeable abyss between humans and other animals, despite resistance[3] .

If apes resemble us too closely for comfort, can we seek solace in an unbridgeable gulf between humans and *artificial* agents? Or is it conceivable that artificial agents might be designed to cope intelligently with their environments in ways that resemble human strategies for living, for example by solving problems creatively, using language, and belonging to co-operative communities that collectively endorse principles of justice and compassion?

In considering this question we shall briefly recapitulate the argument which relies on Gödel's Incompleteness Theorem to substantiate the existence of such an unbridgeable divide. An excursion into research on cognition will lead us to the conclusion that the features of human cognition that are relevant to the main argument are in principle available to embodied artificial agents. Gödel's theorem may then be re-interpreted as demarcating a boundary between disembodied cold cognition and embodied hot cognition, and the gap between embodied artificial agents and humans perceived as a ditch rather than an abyss.

THE CHINESE ROOM THOUGHT EXPERIMENT

The simplest artificial agents are computer programs, like word-processors. You provide input to the program through the keyboard, the program does things to the input, and finally output is delivered to the screen. It is perfectly reasonable to think of the program as an *agent*. And although it is clearly not a very versatile agent, since it cannot pour you a glass of wine to ease the throes of composition, it may often be intelligently helpful, for instance when it corrects your spelling or grammar.

John Searle[4] devised a famous thought experiment to show the magnitude of the difference between such a program and a human being. Imagine a man who does not know any Chinese at all, closed up in a room with, say, a fax machine through which messages can be received and sent. Suppose that enquiries are faxed to the man in Chinese. Although the man is ignorant of Chinese, imagine that there are rules posted up on the walls (the program) which the man can use to decide what mysterious symbols to send in response to the inputs. Is it possible, Searle asks, for it to appear to outside observers as if the man in the room understands Chinese? The man is like a computer program, following rules without understanding them, and the appearance of intelligence is mere appearance lacking in substance.

When we reflect on what a word-processor and Searle's man in the Chinese room have in common, we realise that they share three points of similarity:

- Their only 'sensor' is a communication channel that receives symbols as input.
- They manipulate symbols syntactically – that is, they manipulate symbols by following rules based on the shapes of the symbols, without understanding what the symbols mean.
- Their input is under the control of other agents, so that in effect they are being used as slaves – they lack autonomy.

Let us call such agents *instrumental*.

The most widely known articulation of the claim that humans are separated from artificial agents by an unbridgeable abyss, due to Roger Penrose[5] , takes Gödel's Incompleteness Theorem as establishing limitations of instrumental agents to which humans are not subject. By definition, instrumental agents are limited to computation, since this is just another name for rule-based symbol-manipulation. Human understanding surpasseth computation, by Gödel's theorem. So if we assume that artificial agents are condemned to be instrumental, the conclusion follows.

THE PENROSE ARGUMENT

Consider an instrumental agent like a word-processor or the man in the Chinese room, obliged to process meaningless symbols according to rules based on their shapes. Surprisingly, an agent of this limited sort is able, if the rules are appropriately chosen, to prove theorems of mathematics[6] expressed in the symbols of a formal language. Some impressively difficult and important results in mathematics have been accomplished by such theorem-provers. In 1931, however, Kurt Gödel[7] showed that there exists a sentence G, true of the natural numbers, that cannot be proved by any theorem-prover given as input only a formal description of the properties of the natural numbers (unless the agent also 'proves' a lot of things that aren't true).

The existence of this true-but-unprovable sentence G is Roger Penrose's justification for asserting that the human brain has capabilities that cannot be duplicated by artificial agents such as theorem-proving computer programs. Truth, it is claimed, is a matter of human intuition ("understanding"), a faculty that according to Penrose is so mysterious only (a new form of) quantum mechanics can account for it.

What Penrose seeks from quantum mechanics is nondeterminism at the level of neural synapses. Nondeterminism arises in quantum mechanics because a system can be in a superposition of states, from which the act of measurement wrenches the system unpredictably into one or another particular state. For example[8], imagine that a box is prepared for occupancy by an unfortunate cat (a scenario devised by Erwin Schrödinger, an influential figure in early quantum theory). A radioactive substance like uranium is placed inside the box, together with a detector. A time period is chosen such that there is a 50-50 chance of one of the uranium nuclei decaying within that period, and the detector is set to activate a switch that will release a deadly poison, if such decay occurs. Now the cat is placed in the box and the lid is shut. As the time passes, the system – consisting of the randomly decaying substance, the poison, and the cat, all enclosed in the box – is in a superposition of two states. The cat is simultaneously both dead and alive, and it is only when the lid is opened at the end of the time period, so that a human observer can see what's going on, that the system leaps into one of the two particular states. Only in that instant will the cat instantaneously become either alive or dead, and it is impossible to predict which.

It is such superposition and the consequent nondeterminism that Penrose hopes may explain the power of human intuition to go where no step-by-step theorem-prover dare tread. To evaluate the merits of this claim, we should first improve our grasp on the concept of intuition, which may not be as mysterious as Penrose believes.

What can cognitive science tell us about human intuition? Following Zajonc's 1980 paper[9] a flood of evidence[10] has demonstrated that humans have two information processing systems that work in parallel and that underpin two kinds of cognition, namely reasoning (or *cold* cognition) and intuition (or *hot* cognition).

The following table summarises the differences between the two systems.

Reasoning = cold cognition	**Intuition = hot cognition**
Under conscious control	Automatic
Step-by-step process that is consciously viewable	One-step process in the sense that result suddenly appears in consciousness
Involves syntactic symbol-manipulation	Involves analogical pattern-matching
Demands attentional resources	Does not demand attentional resources
Unique to humans over age 2 and language-trained apes	Common to all mammals
Disembodied and platform-independent (can be performed by any rule-following agent)	Embodied and platform-dependent (depends on the brain and body housing it)
Slow and effortful	Fast and frugal
Cautious and defensive	Blithely uses heuristics
Eschews emotion	Often affectively valenced

Penrose's argument rests upon three claims: that Gödel's theorem places limits on reasoning that do not hold for intuition, that intuition is in principle beyond the capabilities of artificial agents, and that quantum mechanics must be invoked in order to account for the human ability to employ intuition.

The first thing that strikes one upon examining the formulation and proof of Gödel's theorem is that it does not appear to be about intuition in any obvious sense, for the demonstration that the sentence G is true is itself made by a step-by-step argument, in other words via reasoning. The demonstration that G is true of the natural numbers occurs at the meta-level (in this case, the level of set theory), using a more expressive language than that available to the instrumental agent which was attempting to prove theorems about the natural numbers. If we call the symbol-strings manipulated by the agent the object-language, then the meta-language is a language in which we are able to quote all the symbol-strings of the object-language and say things about them that couldn't be said in the object-language itself. Set theory is a more powerful theory than number theory, and a proof that sentence G is true of the natural numbers can be given in set theory, although as Gödel showed no such proof could be given in number theory itself. In a sense, therefore, Gödel's theorem is merely a comparative statement about the relative expressiveness of two language-levels.

However, there is another sense in which Gödel's theorem is indeed about intuition, and this connection depends on the notion of representation.

SEMANTICS

An agent encountering a sensory input (via sensors that resemble human senses such as vision or hearing) forms an iconic representation of the stimulus object. By *iconic* we mean that the representation is an analog of the sensory input (in other words bears a direct relationship – a similarity – to the proximal projection of the distal stimulus object on the agent's transducer surfaces). By a process which is not yet fully understood, but which includes an analog-to-digital conversion, representations which are *symbolic* may be formed. Symbolic representations involve discrete symbols such as those of language, whose association with the signified object is conventional rather than based on similarity. The name of the person whose face we recognise is a symbolic representation of that person, whereas the image we form of the face is an iconic representation.[11]

Reasoning involves symbolic representations (more precisely, the syntactic transformations which change one symbolic representation into another). This syntactic computational process is useful because the symbolic representations are not meaningless[12] . The meaning of a symbolic representation is provided by the iconic representations in which it is grounded, e.g. from which it was manufactured. We say that these grounding iconic representations provide the *semantics* of the symbolic representation. Instrumental agents such as word-processors, theorem-provers, or the man in the Chinese room lack any semantics for the symbolic representations they syntactically manipulate.

The significance of this limitation is that a mathematician demonstrating the Gödel sentence G to be true of the natural numbers is obliged to have recourse to the semantics of G, for truth is a semantic notion operationally defined in terms of denotation. To decide whether a sentence

is true, you have to go and look whether what it says is so, really is so. As it happens, the semantics of mathematical sentences (where sentence = symbolic representation) is abstract, being provided by sets rather than distal stimulus objects living in the physical world. Thus to establish the truth of sentence G, it is sufficient to invoke the language of set theory and to describe how to build up the relevant sets. Consequently the truth of G could be established by reasoning, albeit in a meta-language powerful enough to allow descriptions of the iconic representations (e.g. the set of natural numbers) constituting the semantics of G. But this is unusual. It works only because mathematical objects are abstract.

In the everyday physical world, as distinct from the abstract universe of mathematics, semantics must be found outside language and through the senses. The iconic representations providing the semantics are like pictures, and a picture is worth not a mere thousand words but an infinite number of words. As Harnad[13] puts it:

> Words obviously fall short when they leave out some critical feature that would be necessary to sort some future or potential anomalous instance; but even if one supposes that every critical feature anyone would ever care to mention has been mentioned, a description will always remain essentially incomplete in the following ways:
>
> (a) A description cannot convey the qualitative nature of the object being described (i.e. it cannot yield knowledge by acquaintance), although it can converge on it as closely as the describer's descriptive resources and resourcefulness allow. (Critical here will be the prior repertoire of direct experiences and atomic labels on which the descriptions can draw.)
>
> (b) There will always remain inherent features of the object that will require further discourse to point out; an example would be a scene that one had neglected to mention was composed of a prime number of distinct colours.
>
> (c) In the same vein, there would be all the known and yet-to-be-discovered properties of the prime numbers that one could speak of – all of them entailed by the properties of the picture, all of them candidates (albeit far-fetched ones) for further discourse "about" the picture.
>
> (d) Finally, and most revealingly, there are the inexhaustible shortcomings of words exemplified by all the iterative afterthoughts made possible by, say, negation: for example, "the number of limbs is *not* two ..." The truth of all these potential descriptions is inherent in the picture, yet it is obvious that no exhaustive description would be possible. Hence all descriptions will only approximate a true, complete "description".

Gödel's theorem may now be seen as one instance[14] of a general thesis: that truth, being a semantic notion, ultimately requires the use of iconic representations which, outside of mathematics, can only be approximated by descriptions in language. As the psychologist Lila Gleitman avers, "linguistic systems are *merely* the formal expressive medium that speakers devise to describe their mental representations", and "linguistic categories and structures [serve as] more-or-less straightforward mappings from a pre-existing conceptual space, programmed into our biological nature."[15]

Instrumental agents lack the capacity to ground their symbols semantically since they lack the sensory apparatus with which to gain iconic representations. They are therefore unable to deal with the semantic notion of truth. Intuition is even further beyond them as it involves, not step-by-step reasoning implemented as syntactic manipulation of symbols, but something more like image processing applied to iconic representations. Would equipping an agent with sensors be sufficient to provide it with intuition? To see that this is not the case we must look more closely at the way intuition works, which turns out to involve the capacity to experience emotions.

INTUITION AND THE EMOTIONS

Reflect again on the table which summarises and contrasts the features of reasoning (cold cognition) and intuition (hot cognition). The differences include conscious awareness, and this gives us a way to measure the relative importance of the two forms of cognition. One of the most profound insights to emerge from cognitive science in the past two decades is that most of our everyday thinking and decision-making is performed at a level inaccessible to our conscious awareness[16]. So pervasive is automaticity (the ability of the hot cognitive system to solve problems unconsciously) that perhaps we should pause to consider why its antithesis – reasoning – evolved at all.

John Bargh[17] and Bill Clancey[18] suggest that the purpose of consciousness is to connect a parallel mind to a serial world. One benefit of establishing such a connection is that communication between agents can occur, communication being a sequential process. Communication in turn offers the benefit that agents can learn from one another. Step-by-step reasoning can be cast into the form of serial communication, and is therefore of importance for explanation and social learning. It is also not unknown for an agent to change its mind about something after articulating the evidence for and against, hence reasoning has a role to play as a safety net that helps to catch errors that may be made by hot cognition.

We may conclude that conscious reasoning is a secondary mechanism which supplements hot cognition in agents that have evolved or been designed to form symbolic representations. But how does the primary system of unconscious hot cognition work?

Much of our understanding of hot cognition is due to Antonio Damasio[19]. He studied patients with damage to the ventromedial frontal lobe. Such patients have unimpaired intelligence, as far as IQ tests can determine, but their ability to make sensible decisions is conspicuously reduced. The only deficit exhibited by these patients, and therefore the likely cause of their poor decision-making, is an emotional flatness, a reduced capacity for feeling emotion. Fascinating experiments[20] involving the Iowa Gambling Task shed more light on the relationship between the emotional deficit and the impaired decision-making.

The subject has four decks of cards. Each card contains a reward, but some cards carry a penalty. The subject has an initial stake of $2000. The turning of any card in decks A and B pays a generous $100, while cards in decks C and D pay only $50. However, some cards in decks A and B (the high-paying decks) require the player to make a sudden high payment to the experimenter, sometimes as much as $1250, whereas the cards in decks C and D that carry a penalty impose much smaller sums, typically less than $100. These facts are not disclosed to

the subject, who has to learn by sampling. Normal subjects begin by sampling from all four decks. Lured by the experience of high reward from turning over cards in the A and B decks, they show an early preference for those decks. Gradually, within the first 30 turns, they switch their preference to decks C and D and stick to this strategy to the end. Ventromedial patients, in complete contrast, never change the early preference for the A and B decks. Despite the larger amounts they receive from these cards, the penalties they keep having to pay are so high that halfway through the game they are bankrupt. This is the case even for ventromedial patients who describe themselves as conservative and prudent.

When the experiments were repeated with the subjects hooked up to a polygraph, the profile of skin conductance responses revealed an astonishing pattern. Both normal subjects and ventromedial patients showed skin conductance responses after turning over a card, but quite soon the normal subjects also began to show a skin conductance response while they were deliberating whether to pick a card from the 'bad' decks A and B. Thus the normal subjects were learning to predict a bad outcome – and this occurred well before they were consciously aware of and able to articulate the strategy of avoiding decks A and B. Ventromedial patients, on the other hand, showed no anticipatory responses whatever! Simply put, the ventromedial patients did not learn from their experiences even though, at the time of turning over a card, their experience was emotionally valenced.

Damasio's hypothesis is that the emotional deficit of ventromedial patients undermines their ability to generate and effectively use s*omatic marker*s, neural representations of body states that imbue behavioural options with affective significance and thus guide real-time decision-making. The somatic markers form the basis of the unconscious, automatic, hot cognitive system that we have called 'intuition'. Somatic markers depend on emotions, though not on the conscious awareness of an emotion. There is a difference, Damasio asserts, between an emotion as reflected by changes in the body, on the one hand, and the feeling or conscious awareness of that emotion on the other. To emote is to alter the state of the body – the viscera, the musculature, the internal milieu. Consciousness has to do with continuous signals from the body to the brain that provide an ongoing backdrop, and when these signals result in changes to the brain's map of the body, a conscious feeling is experienced. Indeed, signals from the prefrontal cortex or amygdala can change the body map directly, providing a mechanism of as-if-body and accommodating the roles of imagination and memory in generating feelings.

If intuition depends on somatic markers, which involve emotions, then it makes sense to look more closely at the role of emotions in the decision-making of agents.

EMOTIONS AND EMANCIPATED AGENTS

From the perspective of cognitive science, a human agent may be viewed metaphorically as a cage containing a crowd of inner subagents each of which is functionally specialised for solving problems in a different domain: face recognition, mate choice, heart rate regulation, sleep management, predator vigilance, and so on. The inner subagents are activated by cues from the environment, but there is an organisational challenge to be overcome: some subagents should not be activated together, while others should. For example, sleep and flight from a predator are

mutually inconsistent, and it is no accident that sleep is impossible while your heart is racing with fear. But predator avoidance is enhanced by simultaneous shifts in both heart rate and auditory acuity. There is thus a need for superordinate inner agents that co-ordinate the suppression or activation of subagents. The superordinate inner agents[21] are emotions.

When a condition or situation having a recognisable cue is encountered, a signal is sent out from the superordinate emotion-agent to activate the specific constellation of subagents appropriate for solving the types of adaptive problems associated with that sort of situation. This constellation of subagents is sometimes referred to as the agent's momentary thought-action repertoire. By way of example, consider being alone at night and hearing a sound that indicates the possible presence of a predator. Your momentary thought-action repertoire is affected in the following ways:

- Shifts in perception and attention: you suddenly hear with greater clarity sounds that bear on the hypothesis that you are being stalked but that ordinarily you would not perceive or attend to.
- Shifts in goals and motivational weightings: safety becomes a higher priority than hunger or charming a potential mate.
- Shifts in information-gathering: watching the news becomes less important than checking where the baby is or trying to see whether there is a prowler in the garden.
- Shifts in categorisation: the hallway closet becomes salient as an instance of the category of hiding places rather than the category of storage spaces.
- Shifts in memory retrieval tasks: did the man in the grocery store give you a funny look earlier that day?
- Shifts in communication processes: your face assumes a species-typical fear expression and your voice becomes high-pitched.
- Physiological changes occur: gastric mucosa turn white as blood leaves the digestive tract and heart rate goes up.
- Specialised learning systems are activated: if the threat is real and the ambush occurs, you may experience an amygdala-mediated recalibration that, as in post-traumatic stress disorder, may last for the rest of your life.
- Particular courses of action are potentiated: you resort to fight or to flight or to hiding or even to tonic immobility (paralysis).

Recent work on the positive emotions[22] suggests that the general difference between the positive valence generated by cues from a preferred environment and the negative valence generated by cues from an unsafe environment is that positive emotional states are characterised by a broadened thought-action repertoire, facilitating exploration, learning and creativity[23]; while negative affect is characterised by a narrowing of the repertoire to promote quick and decisive action.

Long before this understanding of human emotion was attained, Herbert Simon[24] suggested that the behaviour of artificial agents might benefit from the influence of an emotional interrupt system, capable of setting aside ongoing programs when real-time needs of high priority are encountered. More recently research has been directed at enabling robots and user-interfaces to recognise human emotions[25].

Although the design of agents having a system of emotion-analogs acting as superordinate inner agents for the co-ordination of specialised subagents is still in its infancy, this does not mean that there is any barrier, either in principle or in practice, to producing simple versions of such designs. It would be a simple matter indeed to equip, say, a robotic explorer with a system of emotional valences based on its remaining fuel supply. While it has plenty of fuel, its affective state is positive and its subprograms for exploration are active. When the fuel supply drops below a critical threshold, its affective state becomes negative and its momentary thought-action repertoire shrinks to focus on returning to the nearest source of fuel.

But can a sufficiently complex system of emotions be designed to provide a substrate for autonomous decision-making in everyday contexts? How complex must a system of emotions be in order for us to recognise a kinship with the artificial agents?

The emotional system must be at least sufficiently complex to afford the capacity for humanlike intuition. Furthermore, we expect agents that are capable of adopting humanlike strategies for coping with their environments both to act autonomously and to engage in collective effort. Emotions are necessary to ensure that such behaviour is responsible and not psychopathic.

This becomes clear as soon as we note that some emotions serve a social function. Self-conscious emotions like embarrassment, shame, guilt, and pride are associated with a process of evaluating our own behaviour and holding ourselves responsible for the effects of our actions[26]. Such emotions regulate the behaviour of individuals in a community and help to render trust and co-operation feasible by providing an internal constraint against cheating. The absence of such emotions would undermine the usefulness of a community as a vehicle for collective problem-solving by permitting system-cheats to flourish without constraint other than external detection.

Psychopaths[27] are the prototypical system-cheats, simulating normality but having no moral compunction about using and abusing other agents to achieve their goals.

There are thought-provoking similarities between psychopaths and the ventromedial patients studied by Damasio. Both categories fail to learn from experience. Both fail to show a skin conductance response to pictures (e.g. of people dying) that arouse strong emotional responses in normal subjects. The causes of their emotional shallowness are different, however, leading to divergence in, for example, language use. Ventromedial patients have no characteristic language deficit. Psychopaths appear to use language in a disjointed fashion – while they are glib and superficially charming, able to speak fluently and to deploy a large vocabulary, to the discerning ear it is apparent that stock phrases are used repeatedly and words and sentences are often incompletely articulated and incoherently strung together[28]. Psychopaths seem to process language syntactically rather than semantically, at least in the sense that they apparently don't 'get' the emotional connotations of words. One is reminded of instrumental agents such as word-processors. In a typical test, the subject watches different words come onto a monitor screen. Some of the words are chosen to have emotional associations, others to be neutral. Whereas normal people respond more quickly to emotional words, psychopaths have the same response time to both emotional and neutral words. When whole brain functional magnetic resonance imaging (fMRI) is used, it is found that normals who process negative emotional words (e.g. rape, death, cancer) display increased activity in the limbic regions of the brain,

whereas for psychopaths there was no increased activity in these regions[29] but over-activation in parts of the brain ordinarily devoted to language, as if emotional material were being analysed in terms of its dictionary definition.

Let us take an emancipated agent to be one that is capable of using humanlike strategies for living, such as co-operation with other agents. An emancipated agent cannot be a psychopath. Co-operation with other members of a community is incompatible with being a system-cheat. Psychopaths are system-cheats for a number of reasons. Not only are they relatively shielded from the self-conscious emotions like shame but their emotional shallowness also blocks the formation of moral sentiments. In normal human beings, beliefs can undergo a process of change from being a matter of preference ("I don't eat meat because I feel healthier without it") to being a matter of moral conviction ("I don't eat meat because killing animals is wrong")[30]. Moral convictions are entrenched – they change not in response to reasoned arguments but only to align with the beliefs of other members of the community[31]. The entrenched status of moral convictions has evolutionary advantages for a community of co-operating agents, for the permanence of such beliefs renders agents predictable and eliminates the concern that fellow agents may be engaged in daily calculations of the profits or costs of actions. The entrenchment appears to be accomplished by the recruitment of emotions such as disgust. Without a system of emotions that is properly functioning and sufficiently broad to provide a means for the formation of moral sentiments, autonomous artificial agents could not be trusted any more than psychopaths can. If we are to avoid the pitfalls illustrated by such celluloid fantasies as "The Matrix" and "I, Robot", our artificial agents had better possess a full panoply of emotions.

CONCLUSION

Penrose is correct in construing Gödel's theorem as distinguishing between the abilities of instrumental agents, which are limited to cold cognition, and humans, who have intuition. However, Penrose is mistaken in relying on this result to accord humans a special status, for instrumental agents such as theorem-provers do not constitute the theoretical pinnacle of achievement for designers of artificial agents.

Instrumental agents may be contrasted with emancipated agents – autonomous agents which are embodied, affective, intuitive, and conscious of their feelings. Embodiment implies possession of their own sensors/senses through which to gain iconic representations. Affect implies possession of organisational 'superordinate' agents working at the level of iconic representations to activate sets of subagents, this being the key to a system of emotions. Intuition implies possession of ways to transform iconic to symbolic representations that apply affective tags to the symbolic representations (somatic markers). The conscious feeling of emotions is closely associated with the ability of an agent to monitor its own body states, which is valuable not only from a standpoint of individual survival but also because it is a prerequisite for the ability of an agent to articulate its emotional state, which in turn facilitates social learning. Finally, the incorporation of a suitably broad system of emotions into agent design provides the means for the entrenchment of moral sentiments, ensuring fitness for participating in communities.

As we improve our understanding of what the design of emancipated agents would require, human intuition becomes less mysterious. We may therefore reject Penrose's final claim also. If ordinary biology in the form of the somatic marker hypothesis can account for intuition, then it is clearly unnecessary to call upon quantum mechanics for the purpose. Schrödinger's cat need not be let into the Chinese room.

1 FBM de Waal, *The Ape and the Sushi Master: Cultural Reflections of a Primatologist* (New York: Basic Books, 2001).

2 S Savage-Rumbaugh and R Lewin, *Kanzi: The Ape at the Brink of the Human Mind* (New York: Wiley, 1994).

3 J Marks, *What it Means to be 98% Chimpanzee: Apes, People, and their Genes* (Berkeley: University of California Press, 2002).

4 JR Searle, "Is the brain's mind a computer program?" *Scientific American* 262, no. 1 (1990): 26-31. See also online: http://plato.stanford.edu/entries/chinese-room/

5 R Penrose, *The Emperor's New Mind* (Oxford: Oxford University Press, 1989); R Penrose, *Shadows of the Mind* (Oxford: Oxford University Press, 1995); R Penrose, *The Large, the Small, and the Human Mind* (Cambridge: Cambridge University Press, 1997).

6 L Wos and GW Pieper, *A Fascinating Country in the World of Computing: Your Guide to Automated Reasoning* (Singapore: World Scientific Publishing, 1999).

7 K Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I" *Monatshefte für Mathematik und Physik* 38 (1931): 173-198.

8 DR Hofstadter, "Heisenberg's uncertainty principle and the many-worlds interpretation of quantum mechanics" *Metamagical Themas* (New York: Basic Books 1985), 455-477.

9 RB Zajonc, "Feeling and thinking: Preferences need no inferences" *American Psychologist* 35 (1980): 151-175.

10 J Haidt, "The emotional dog and its rational tail: A social intuitionist approach to moral judgment" *Psychological Review* 108, no. 4 (2001): 814-834 (see page 818 in particular).

11 T Deacon, *The Symbolic Species: The Co-evolution of Language and the Human Brain* (London: Penguin, 1997), 70.

12 S Harnad, "Computation is just interpretable symbol manipulation, cognition isn't" *Minds and Machines* 4, (1995): 379-390.

13 S Harnad, "Category induction and representation" in Harnad S (ed): *Categorical Perception: The Groundwork of Cognition* (Cambridge: Cambridge University Press, 1987), 535-565.

14 This is a surprisingly weak instance. Since the meaning of a mathematical sentence like G may be sought in set theory rather than in iconic representations of physical objects, there is no obstacle in principle to the design of an instrumental theorem-proving agent capable of manipulating symbols at both the object- and the meta-levels and thus establishing the truth of G purely 'by reason'. But then a new sentence G' can be constructed (in the meta-language, set theory) which is unprovable in set theory, but true in meta-set-theory (i.e. in some model of set theory). And so forth. However, every G, G', ... is provable in some formal system, so none is unprovable in an absolute sense.

15 P Li and L Gleitman, "Turning the tables: Language and spatial reasoning" *Cognition* 83 no. 3 (2002): 265-294, quoted in G Marcus, *The Birth of the Mind* (New York: Basic Books, 2004), 125.

16 JA Bargh and TL Chartrand, "The unbearable automaticity of being" *American Psychologist* 54 (1999): 462-479.

17 Quoted in DG Myers, *Intuition*, (New Haven: Yale University Press, 2002), 29.

18 WJ Clancey, *Conceptual Coordination: How the Mind Orders Experience in Time* (Mahwah: Lawrence Erlbaum, 1999).

19 AR Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (New York: Putnam, 1994); AR Damasio, *The Feeling of What Happens* (San Diego: Harcourt, 1999); AR Damasio, *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain* (San Diego: Harcourt, 2003).

20 A Bechara, D Tranel, H Damasio and AR Damasio, "Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex" *Cerebral Cortex* 6 (1996): 215-225; A Bechara, H Damasio, D Tranel and AR Damasio, "Deciding advantageously before knowing the advantageous strategy" *Science* 275 (1997): 1293-1295.

21 L Cosmides and J Tooby, "Evolutionary psychology and the emotions" in M Lewis & JM Haviland-Jones (eds), *Handbook of Emotions,* 2nd edition, (New York: Guilford Press, 2000), 91-115.

22 B Fredrickson, "The role of positive emotions in Positive Psychology: The broaden-and-build theory of positive emotions" *American Psychologist* 56 (2001): 218-226.

23 AM Isen and B Means, "The influence of positive affect on decision-making strategy" *Social Cognition* 2 (1983): 18-31; AM Isen, KA Daubman and G Nowicki, "Positive affect facilitates creative problem-solving" *Journal of Personality and Social Psychology* 52 (1987): 1122-1131.

24 H Simon, "Motivational and emotional controls of cognition" *Psychological Review* 74 no. 1 (1967): 29-39.

25 R Picard, *Affective Computing* (Cambridge MA: MIT Press, 1998).

26 M Lewis, "Self-conscious emotions: embarrassment, pride, shame, and guilt" in M Lewis & JM Haviland-Jones (eds), *Handbook of Emotions,* 2nd edition, (New York: Guilford Press, 2000), 623-635.

27 RD Hare, *Without Conscience* (New York: Guilford Press, 1998).

28 SE Williamson, *Cohesion and Coherence in the Speech of Psychopathic Criminals* (unpublished Ph.D Thesis, University of British Columbia, 1991), cited in CA Brinkley, A Bernstein and JP Newman, "Coherence in the narratives of psychopathic and nonpsychopathic criminal offenders" *Personality and Individual Differences* 27, no. 3 (1999): 519-530.

29 KA Kiehl, AM Smith, RD Hare, A Mendrek, BB Forster, J Brink and PF Liddle, "Limbic abnormalities in affective processing by criminal psychopaths as revealed by functional magnetic resonance imaging" *Biological Psychiatry* 50 (2001): 677-684.

30 P Rozin, M Markwith and C Stoess, "Moralization and becoming a vegetarian: The transformation of preferences into values and the recruitment of disgust" *Psychological Science* 8 (1997): 67-73.

31 J Haidt, "The emotional dog gets mistaken for a possum" *Review of General Psychology*, to appear (2004).

Johannes Heidema is professor of Mathematics in the Department of Mathematical Sciences at the University of South Africa. His research interests include the history and philosophy of Logic and Mathematics and of science in general. He works with Willem Labuschagne on Applied Logic, and in particular on the nature of cognitive agents and their reasoning with defeasible logics for generating and changing knowledge and beliefs.

Willem Labuschagne is a senior lecturer in the Department of Computer Science at the University of Otago in Dunedin, New Zealand and a member of its Artificial Intelligence Research Laboratory. His research is devoted to the exploration of the logical semantics of nonmonotonic logic. His current goal is to find psychological and epistemological foundations for nonmonotonic logic, a task enriched by the perspective that Johannes Heidema contributes.