The role of the self process in embodied machine consciousness

Owen Holland

What is machine consciousness?

A new area of research dedicated to the construction of machines that are conscious like you – really conscious, not just mimicking consciousness, and with real feelings, not just simulated feelings.

What is machine consciousness?

A new area of research dedicated to the construction of machines that are conscious like you – really conscious, not just mimicking consciousness, and with real feelings, not just simulated feelings.

This is STRONG machine consciousness.

WEAK machine consciousness is aimed at machines that merely simulate consciousness, or that only deal with forms of cognition associated with consciousness.

‘Consciousness is a peculiar phenomenon. It is riddled with deceit and self-deception; there can be consciousness of something we were sure had been erased by an anaesthetic; the conscious *I* is happy to lie up hill and down dale to achieve a rational explanation for what the body is up to; sensory perception is the result of a devious relocation of sensory input in time; when the consciousness thinks it determines to act, the brain is already working on it; there appears to be more than one version of consciousness present in the brain; our conscious awareness contains almost no information but is perceived as if it were vastly rich in information. ***Consciousness is peculiar***.’

Tor Norretranders ‘The User Illusion’ 1991 (tr 1998)

How did consciousness arise?

How did consciousness arise?

We don’t know (just as we don’t know exactly what consciousness ***is***) but it was probably something to do with the development of high intelligence

How did consciousness arise?

We don’t know (just as we don’t know exactly what consciousness ***is***) but it was probably something to do with the development of high intelligence

How did intelligence arise?

How did consciousness arise?

We don’t know (just as we don’t know exactly what consciousness ***is***) but it was probably something to do with the development of high intelligence

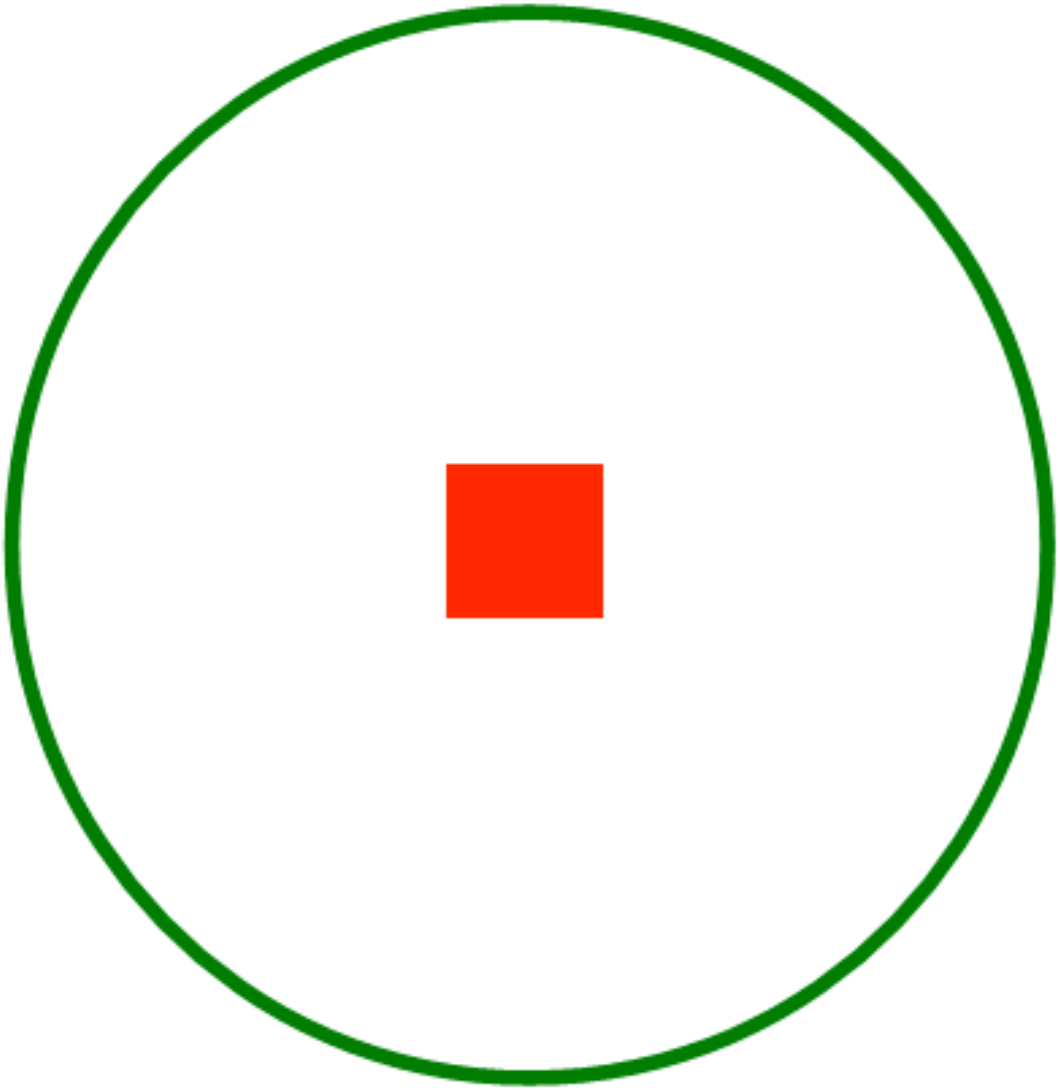
How did intelligence arise?

Through natural and sexual selection – and we can almost understand how and why.

Let’s think about what intelligence has to do….

Let’s consider the problems of an autonomous embodied agent (an animal or robot)…

Let’s consider the problems of an autonomous embodied agent (an animal or robot) in a complex, occasionally novel, dynamic, and hostile world…



Let’s consider the problems of an autonomous embodied agent (an animal or robot) in a complex, occasionally novel, dynamic, and hostile world, in which it has to achieve some task (or mission).

How could the agent ***best*** achieve its task?

- by being preprogrammed for every possible contingency? No
- by having learned the consequences for the achievement of the mission of every possible action in every contingency? No
- by having **learned enough** to be able to **predict the consequences** of tried and untried actions, by being able to **evaluate** those consequences for their likely **contribution to the mission**, and by **selecting** a relatively **good** course of action? Maybe…

But how could it predict?

But how could it predict?

For actions it has tried before in these circumstances, it could simply remember what happened last time

But how could it predict?

For actions it has tried before in these circumstances, it could simply remember what happened last time

If things are only slightly different, it could simply generalise from what it has learned

But how could it predict?

For actions it has tried before in these circumstances, it could simply remember what happened last time

If things are only slightly different, it could simply generalise from what it has learned

But best of all, it could run some kind of ***simulation*** of its potential actions in the world, enabling it to predict their effects – ***even if they involve novel situations or actions***

Here’s how Richard Dawkins puts it:

“Survival machines that can simulate the future are one jump ahead of survival machines who can only learn on the basis of overt trial and error.”

Dawkins, 1976, The Selfish Gene

Two questions:

What exactly has to be simulated?

What is needed for simulation?

What exactly has to be simulated?

Whatever affects the mission. In an ***embodied*** agent, the agent can only affect the world through the actions of its body in and on the world, and the world can only affect the mission by affecting the agent's body.

So it needs to simulate those aspects of its **BODY** that affect the world in ways that affect the mission, along with those aspects of the **WORLD** that affect the body in ways that affect the mission.

What is needed for simulation?

Some structures or processes corresponding to states of the world that, when operated on by processes or structures corresponding to actions, yields outcomes corresponding to the consequences of those actions.

I like to call these structures or processes ‘internal models’, because they are like Kenneth Craik’s ‘working models of reality’ rather than images or static representations

What is needed for simulation?

So we require a model (or linked set of models) that includes the body, and how it is controlled, and the spatial aspects of the world, and the (kinds of) objects in the world, and their spatial arrangement.
But consider…

What is needed for simulation?

The ***body*** is always present and available, and changes slowly, if at all. When it moves, it is usually because it has been commanded to move.

What is needed for simulation?

The ***body*** is always present and available, and changes slowly, if at all. When it moves, it is usually because it has been commanded to move.

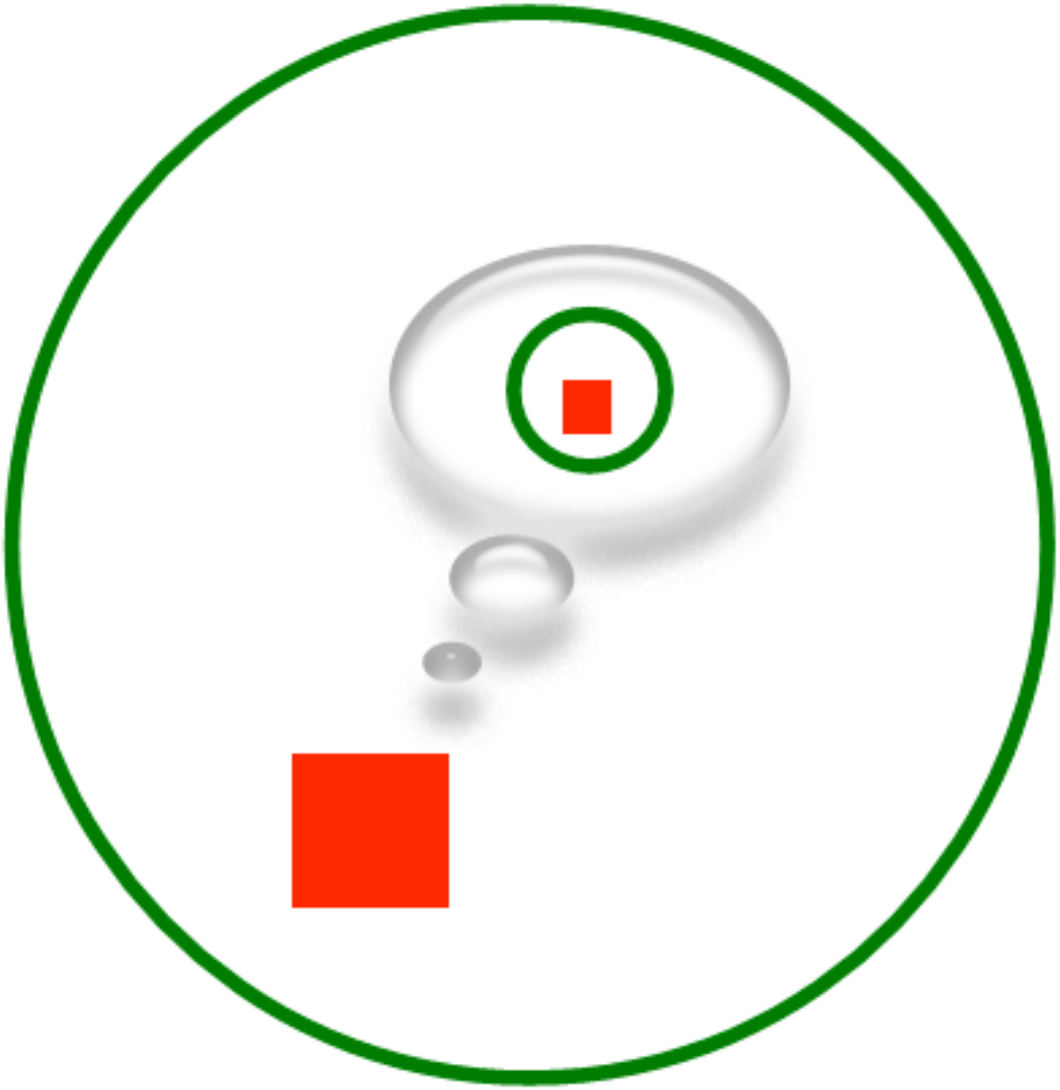
The ***world*** is different. It is ‘complex, occasionally novel, dynamic, and hostile’. It’s only locally available, and may contain objects of known and unknown kinds in known and unknown places.

What is needed for simulation?

The ***body*** is always present and available, and changes slowly, if at all. When it moves, it is usually because it has been commanded to move.

The ***world*** is different. It is ‘complex, occasionally novel, dynamic, and hostile’. It’s only locally available, and may contain objects of known and unknown kinds in known and unknown places.

How should all this be modelled? As a single model containing both body and world?

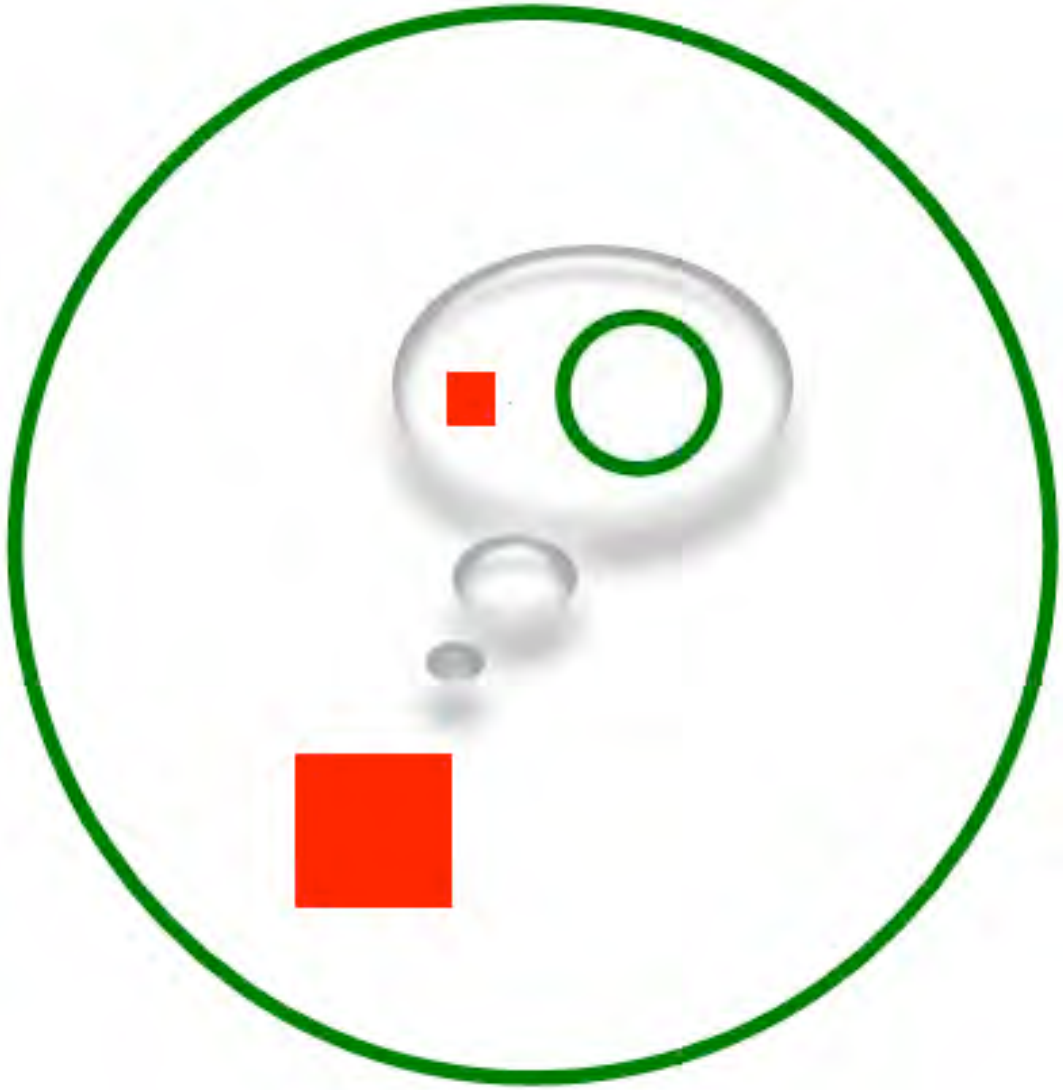


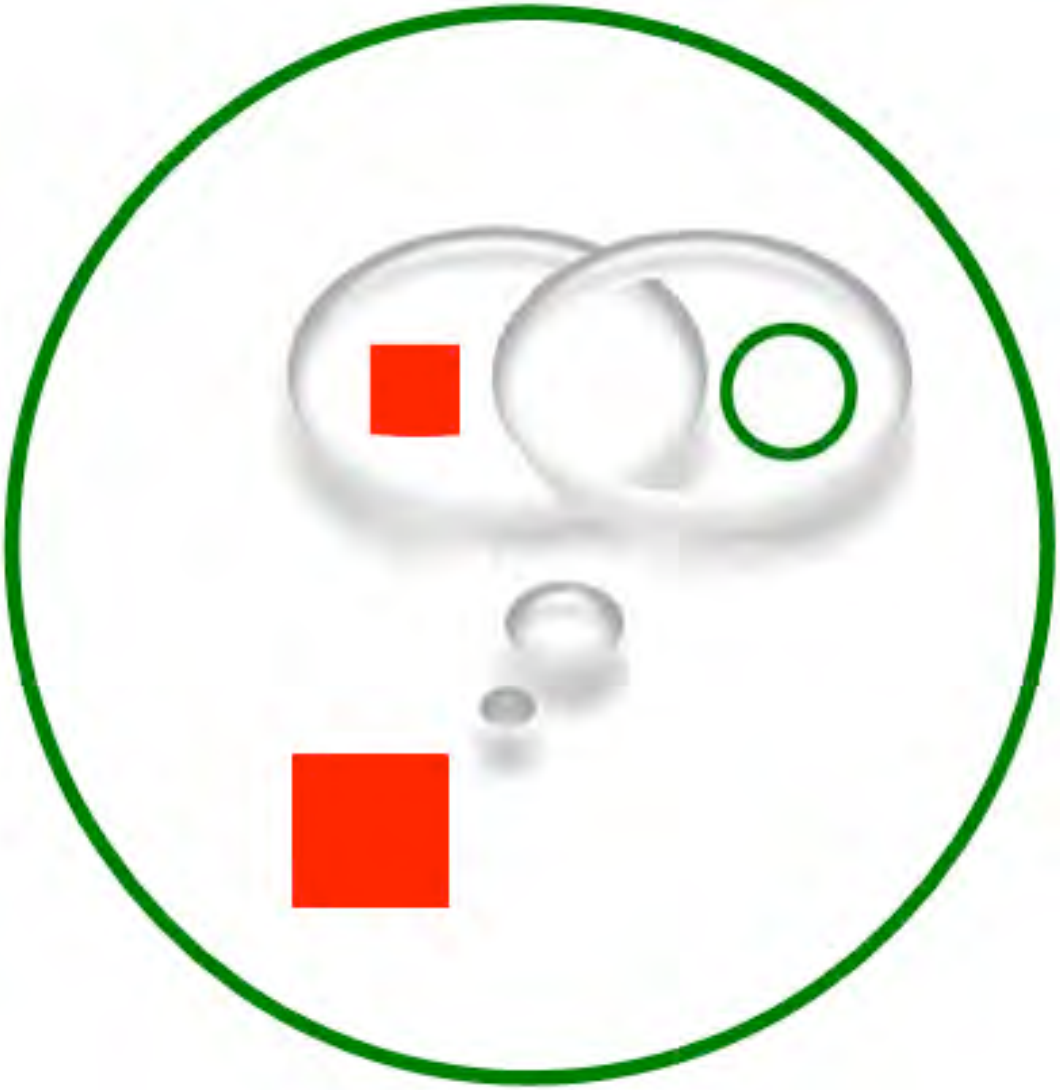
What is needed for simulation?

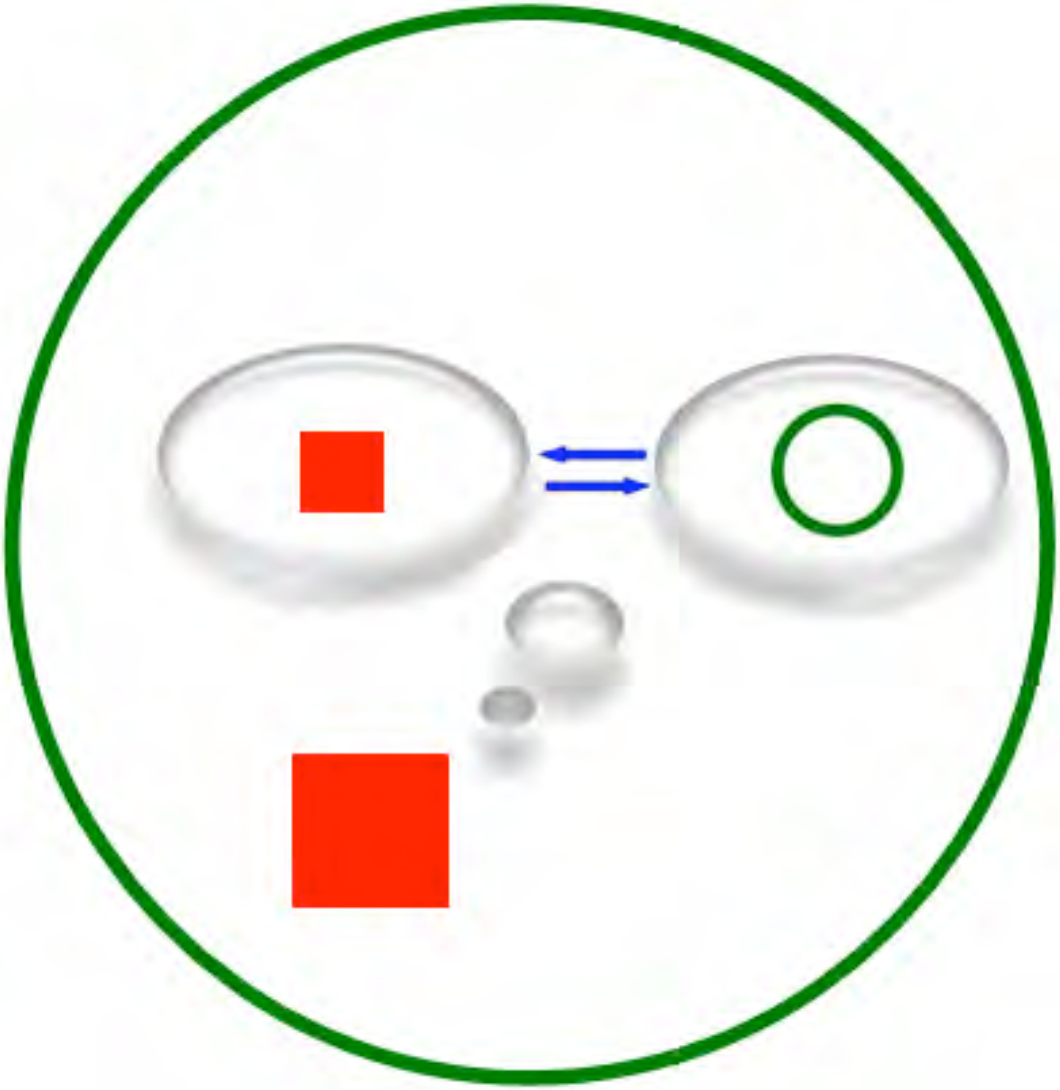
The ***body*** is always present and available, and changes slowly, if at all. When it moves, it is usually because it has been commanded to move.

The ***world*** is different. It is ‘complex, occasionally novel, dynamic, and hostile’. It’s only locally available, and may contain objects of known and unknown kinds in known and unknown places.

How should all this be modelled? As a single model containing both body and world? ***Or as a separate model of the body coupled to and interacting with a separate model of the world?***







Does the (human) brain model the body?

Yes, in many ways. It models the muscular control of movement. It also predicts the nature and timing of the internal and external sensory inputs that will be produced if the movement is executed correctly.

Does the (human) brain model the body?

Yes, in many ways. It models the muscular control of movement. It also predicts the nature and timing of the internal and external sensory inputs that will be produced if the movement is executed correctly.

Ramachandran and Blakeslee describe a host of body image phenomena involving phantom limbs. In one case, a patient with congenital absence of both arms had apparently ‘normal’ phantom limbs from an early age. Some components of the internal model of the body may be inborn.

Does the (human) brain model the body?

Yes, in many ways. It models the muscular control of movement. It also predicts the nature and timing of the internal and external sensory inputs that will be produced if the movement is executed correctly.

Ramachandran and Blakeslee describe a host of body image phenomena involving phantom limbs. In one case, a patient with congenital absence of both arms had apparently ‘normal’ phantom limbs from an early age. Some components of the internal model of the body may be inborn.

But some people have very unusual body plans, yet manage perfectly well…

Myrtle
Corbin
b 1868

Myrtle
Corbin
b 1868

The Commerce Journal.

COMMERCE, HUNT COUNTY, TEXAS, FRIDAY, JULY 22, 1910.

WOMAN WITH FOUR LEGS.

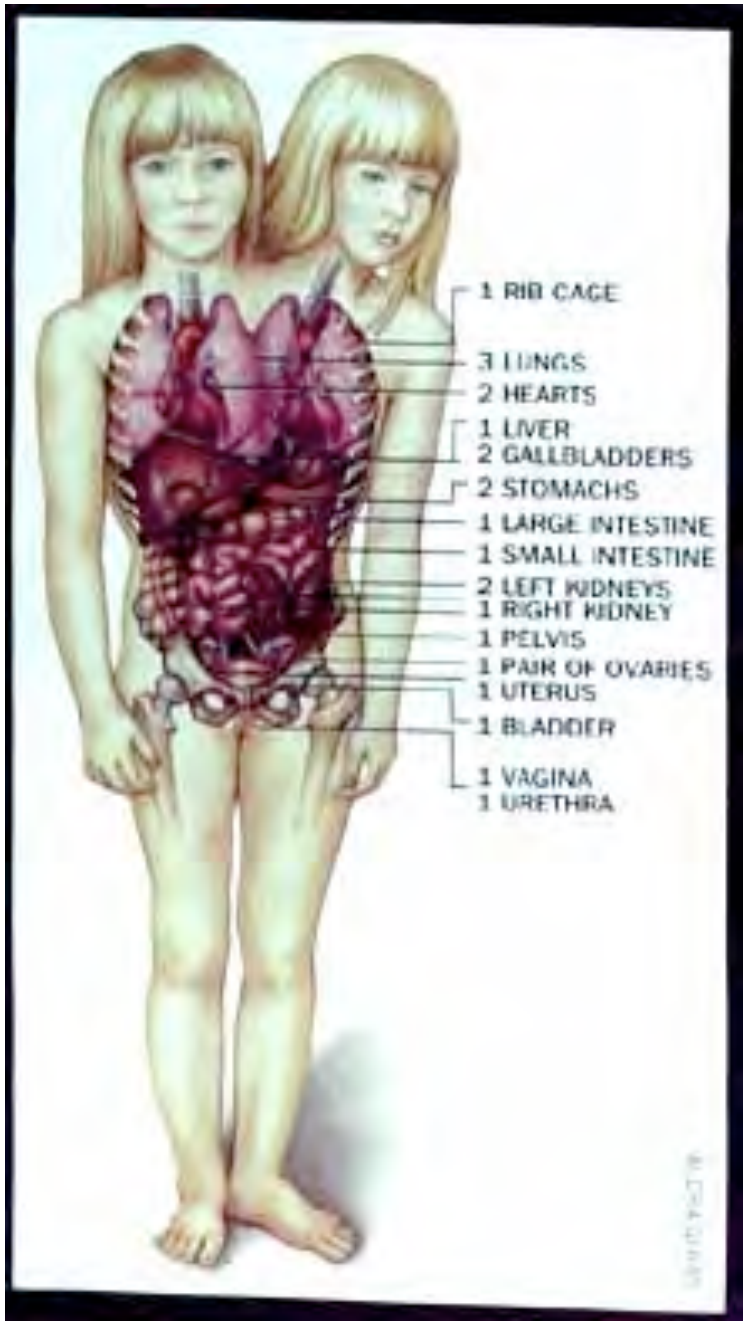
A World Wonder Now Visiting Relatives Near Commerce, Was At Picnic.

A most unusual woman has been visiting relatives near this city for several weeks and, with her husband, attending picnics, at which places she is on exhibition in a tent. In brief, she

One of the surprising things about Mrs. Bicknell is that while she has four limbs, as shown in the accompanying picture, the two inside ones are smaller than the others, thus enabling her to get about with no inconvenience and without any one suspecting her true condition. Her youngest child is a pretty, red cheeked baby of some two years, and like its brothers and sisters, is perfectly developed.

Abigail and Brittany Hensel

Abigail and Brittany Hensel



Does the (human) brain model the world?

Yes, in many ways. It models space, and it models the nature and behaviour of objects, and much of this modelling is innate.

Useful reading (for me anyway):

Wild Minds, by Marc Hauser.

Folk Physics for Apes, by Daniel Povinelli

What has this to do with consciousness?

What Dawkins (1976) said next:

"Survival machines that can simulate the future are one jump ahead of survival machines who can only learn on the basis of overt trial and error...***The evolution of the capacity to simulate seems to have culminated in subjective consciousness…Perhaps consciousness arises when the brain's simulation of the world becomes so complete that it must include a model of itself."***

In other words…

Intelligence may depend on the possession and manipulation of an internal model of the agent (the IAM) interacting with an internal model of the world

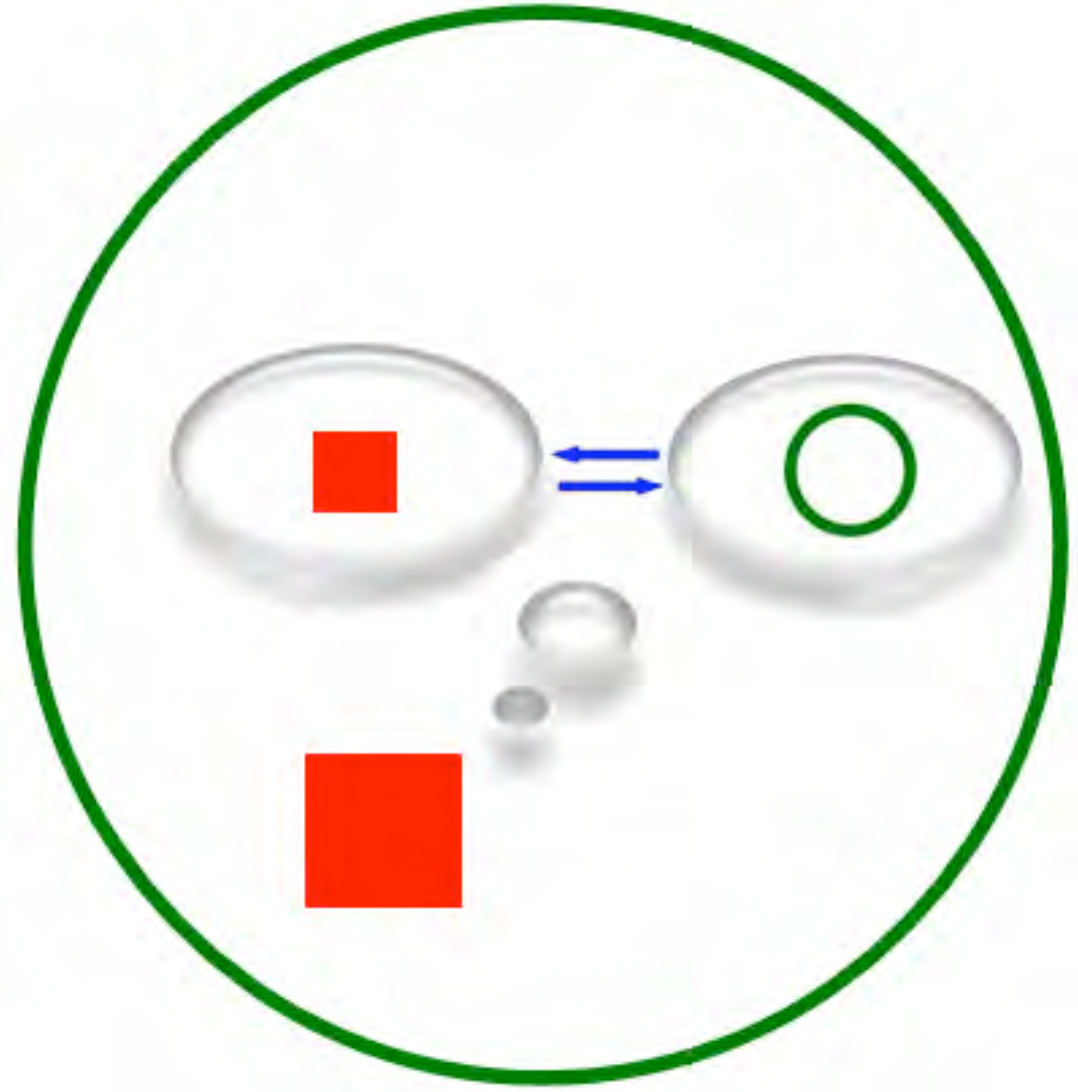
AND

the presence and interaction of these models may also underlie the production of consciousness.

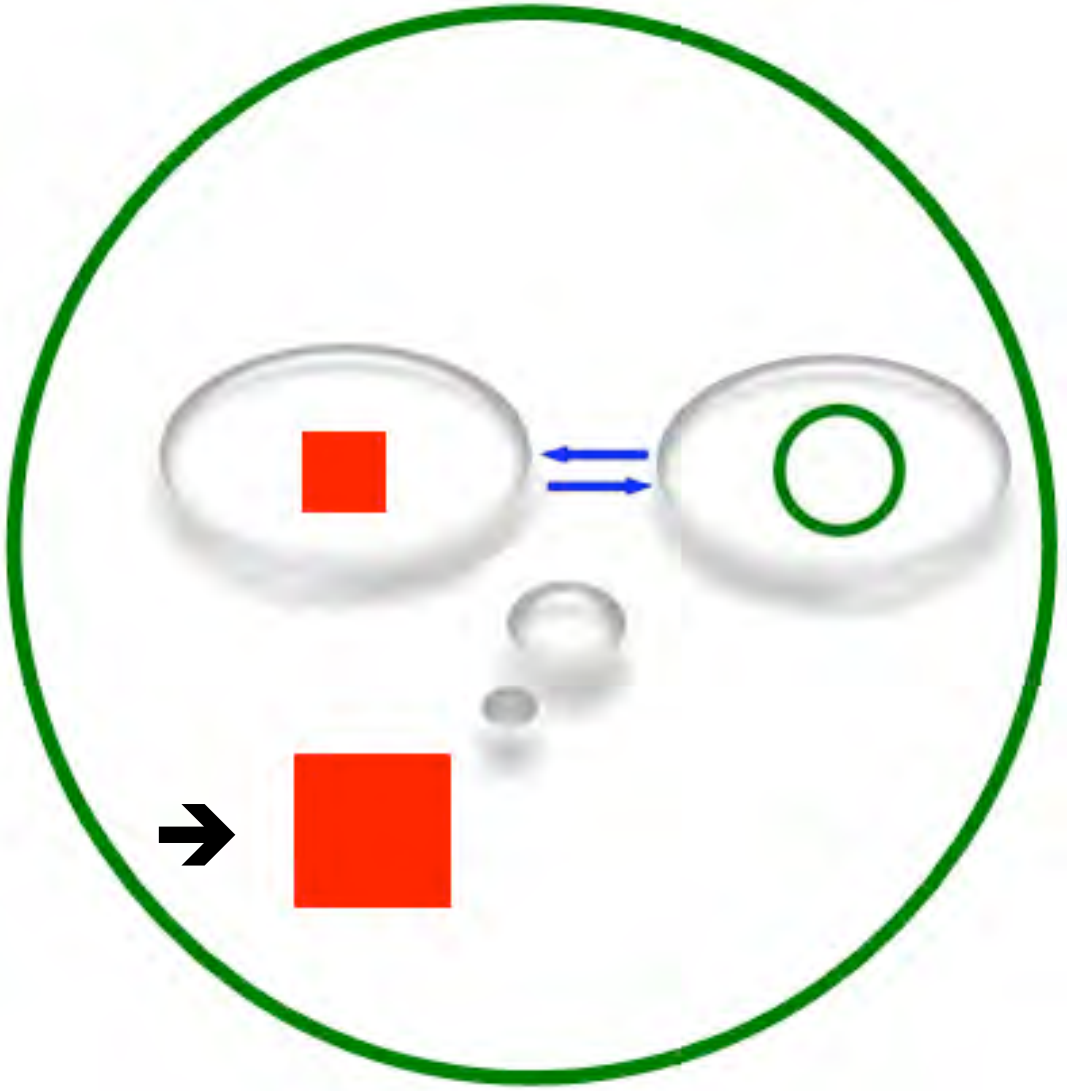
Back to our agent…

Consider things from the point of view of the internal agent model (IAM).

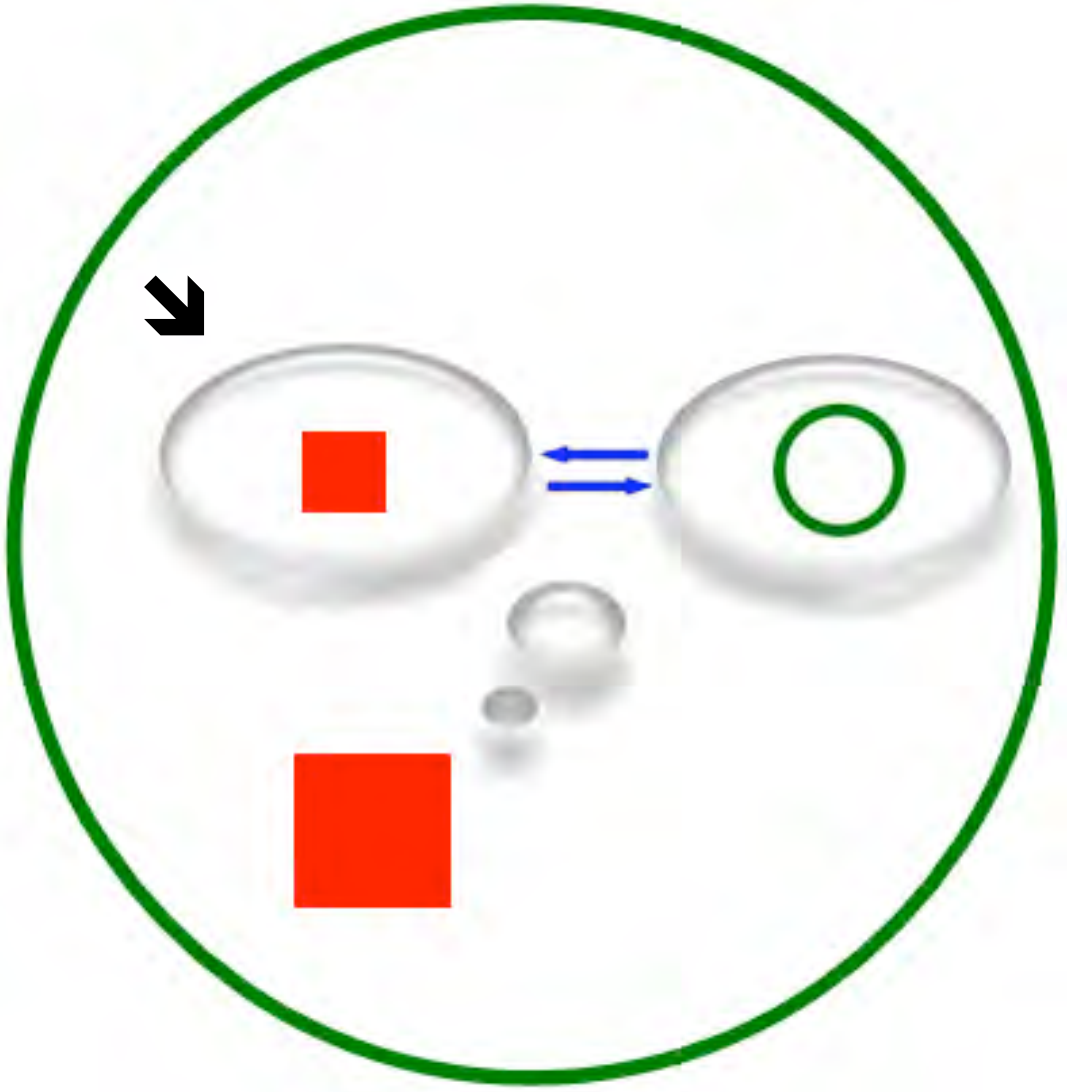
Suppose the model and the simulations are really good…



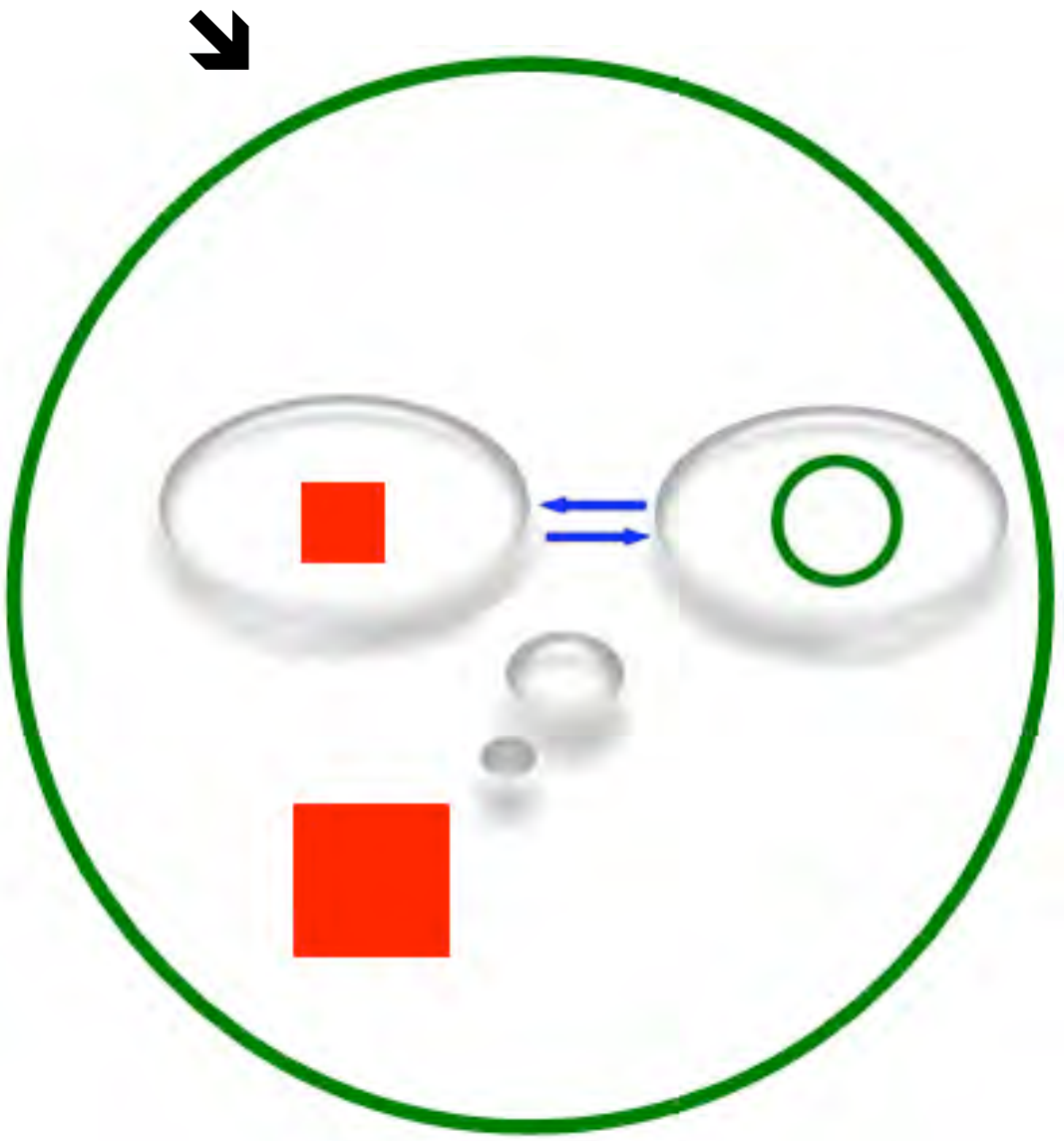
It (the model) will ‘think’ it’s this, the agent…



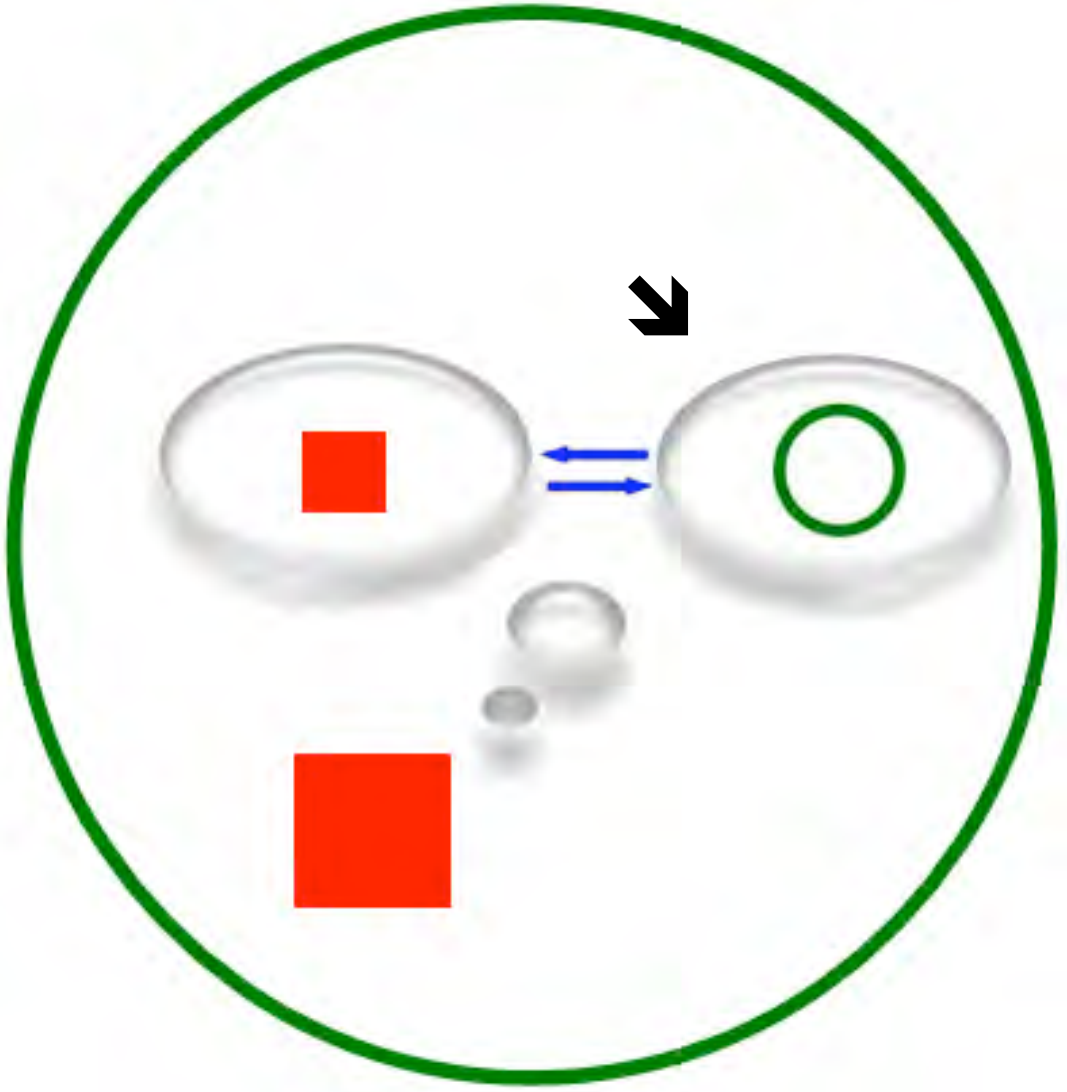
…but it's actually this – a model of the agent



It will ‘think’ it’s interacting with the real world



But it’s actually interacting with a model of the real world



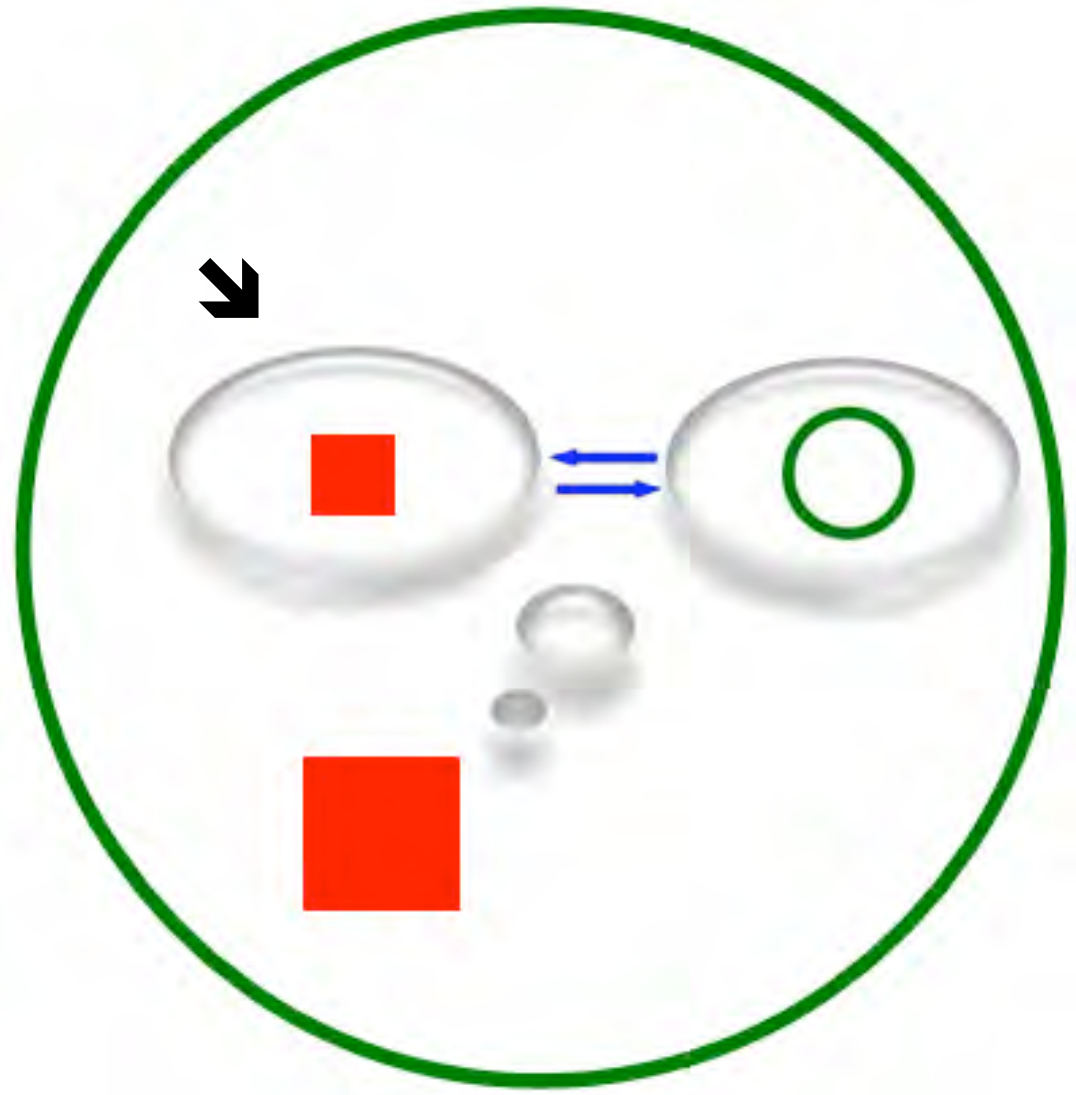
Blackmore's hypothesis

In the late 1980s, the psychologist Sue Blackmore proposed that the conscious self – the experiencing entity – was in fact an internal self model.

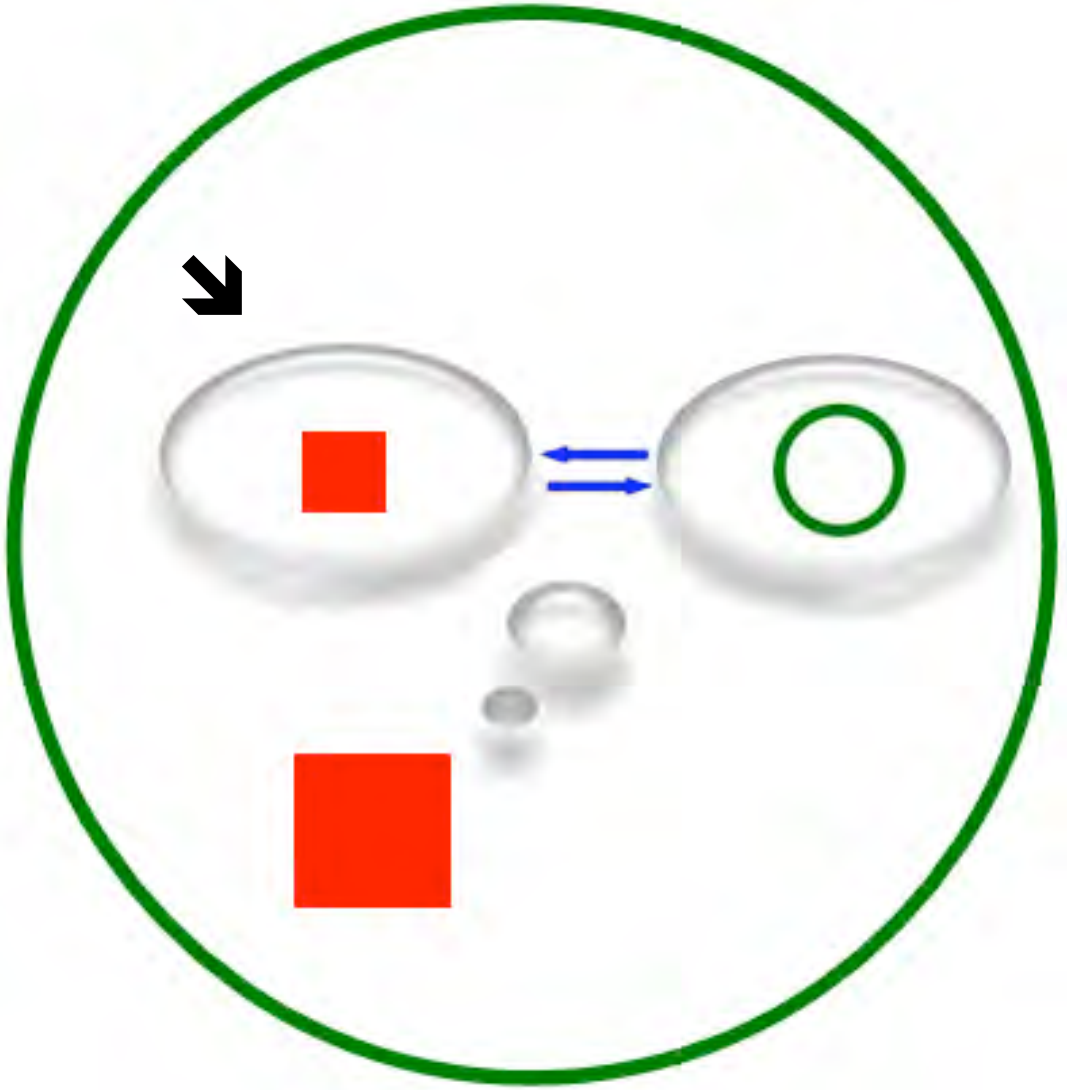
No-one took the idea seriously at the time, but it has surfaced again more recently, notably in the work of Thomas Metzinger.

If it's true…

Consciousness and feelings are in the Internal Agent Model (the IAM) – the system's 'software' model of itself…

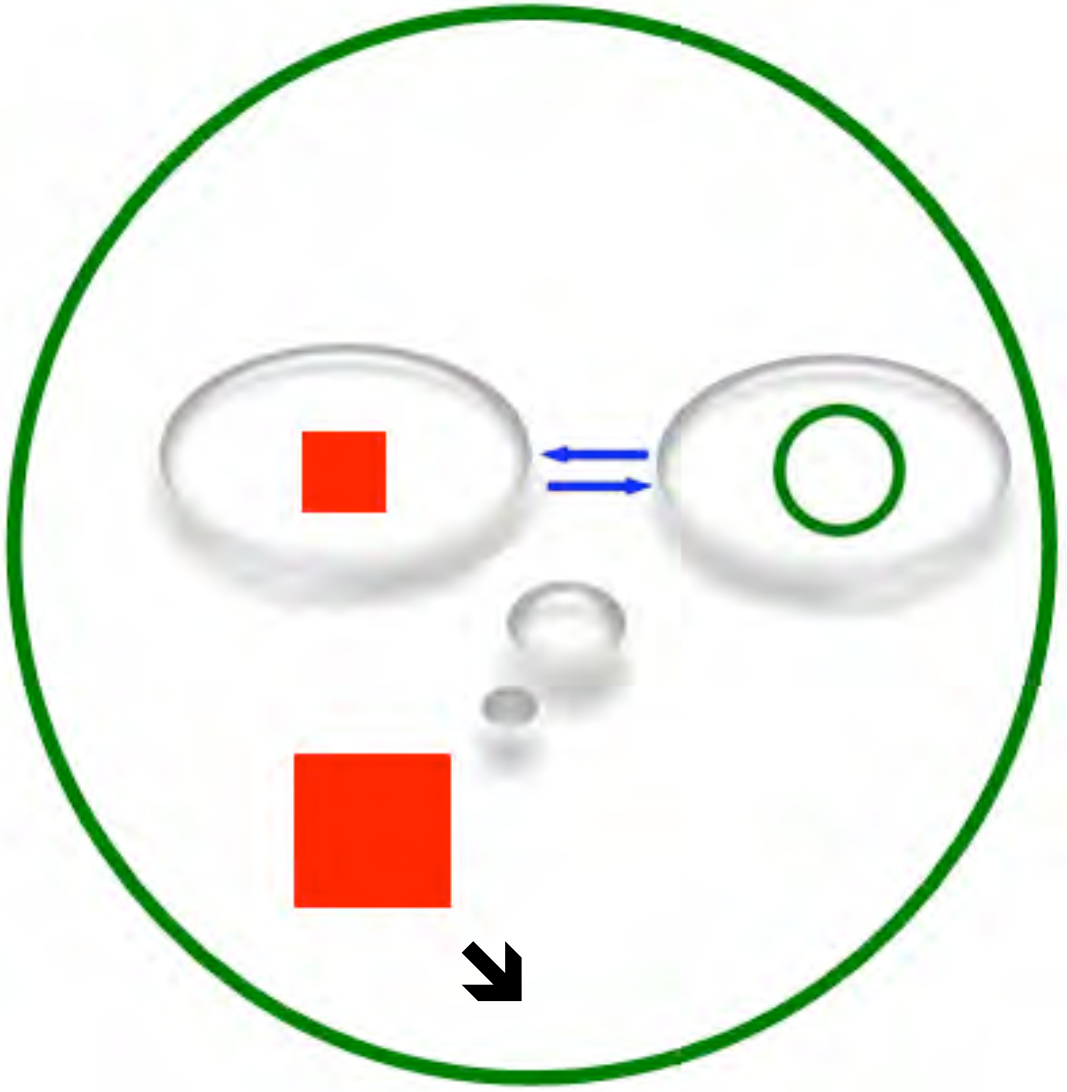


Consciousness and feelings are in the Internal Agent Model (the IAM) – the system's 'software' model of itself…



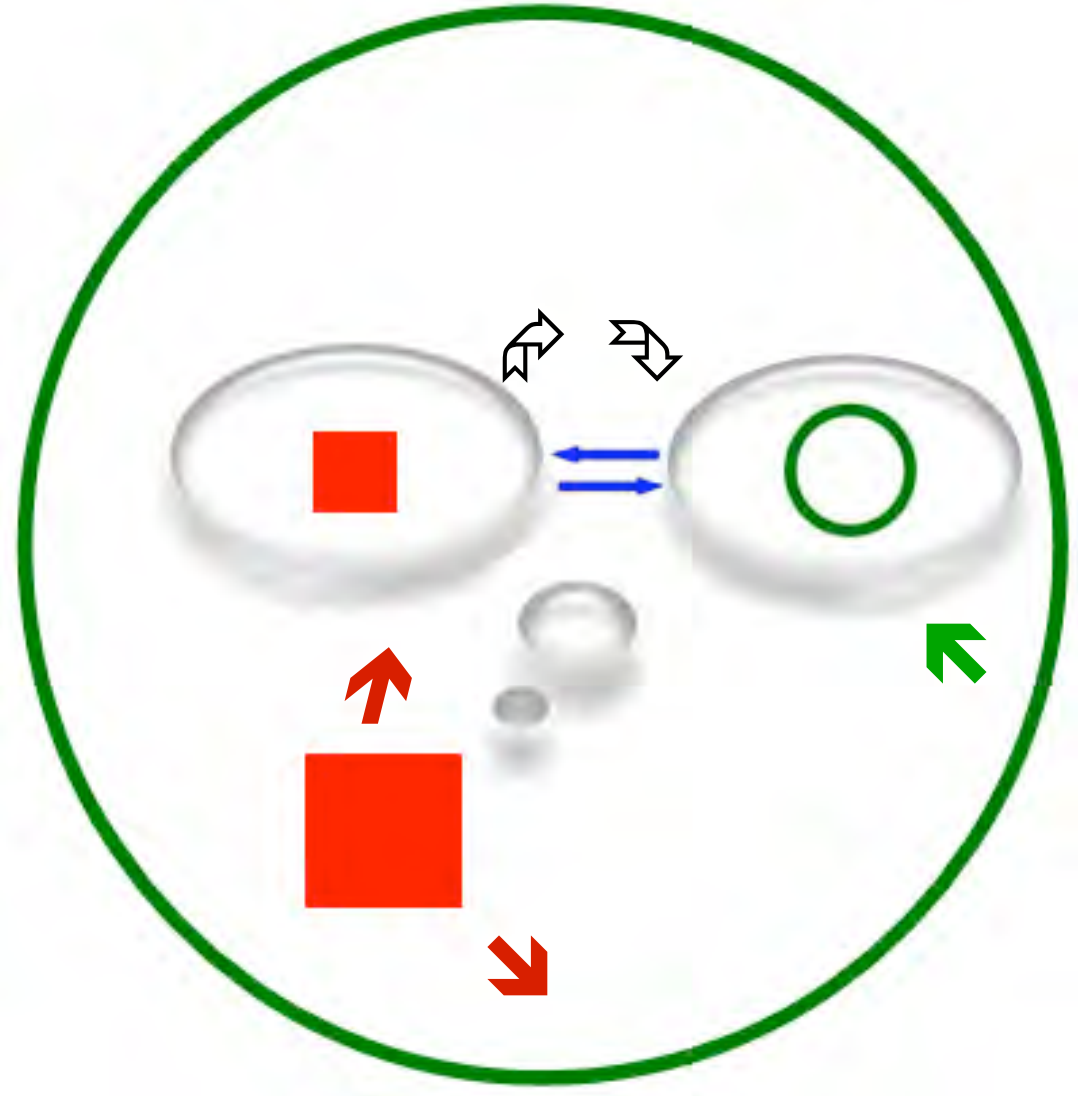
…and feelings are what influence the evaluative function, enabling the choice of 'good' actions.

You think you control your body, and act on the real world

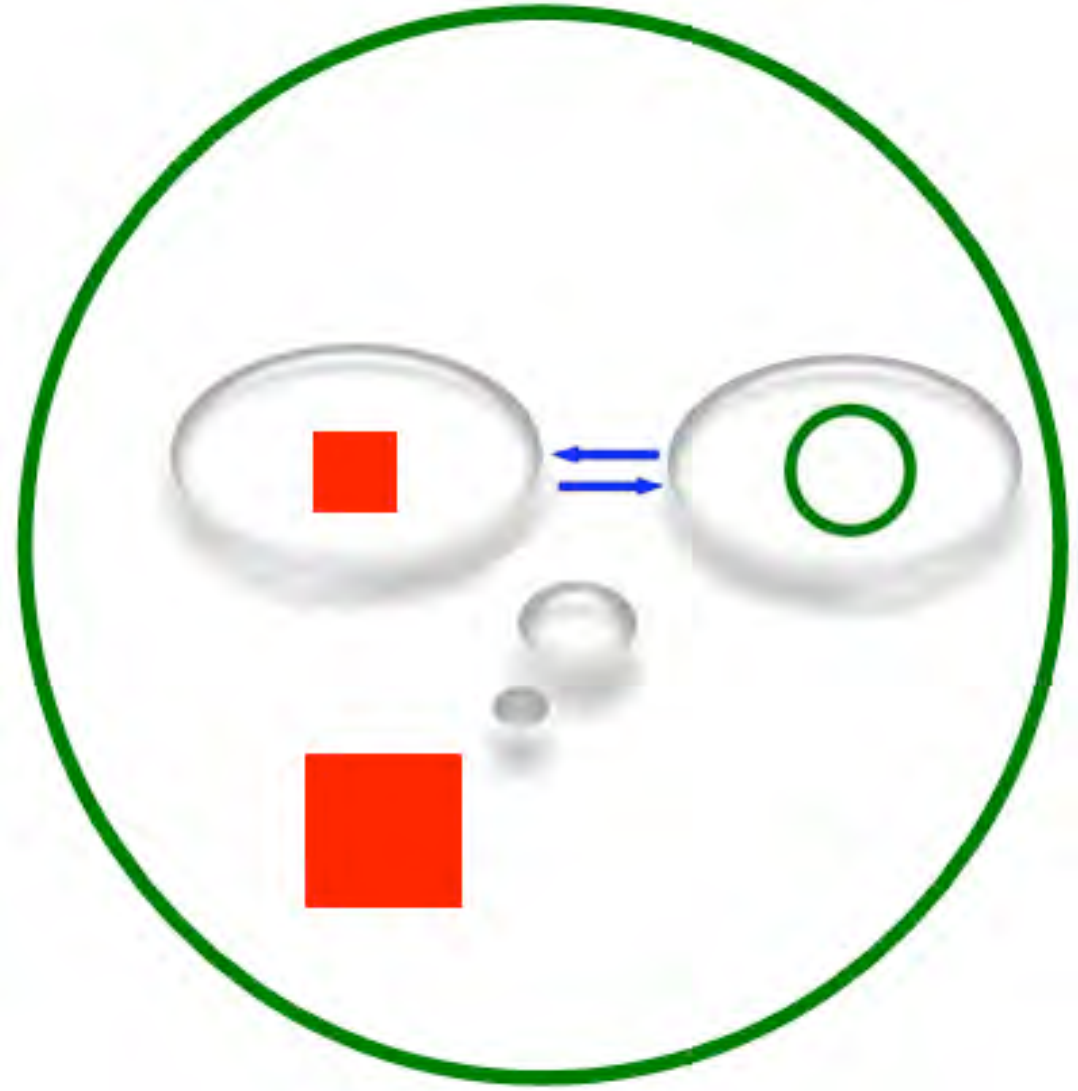


But your body is controlled by other structures within your brain, using the information about ‘good’ choices.

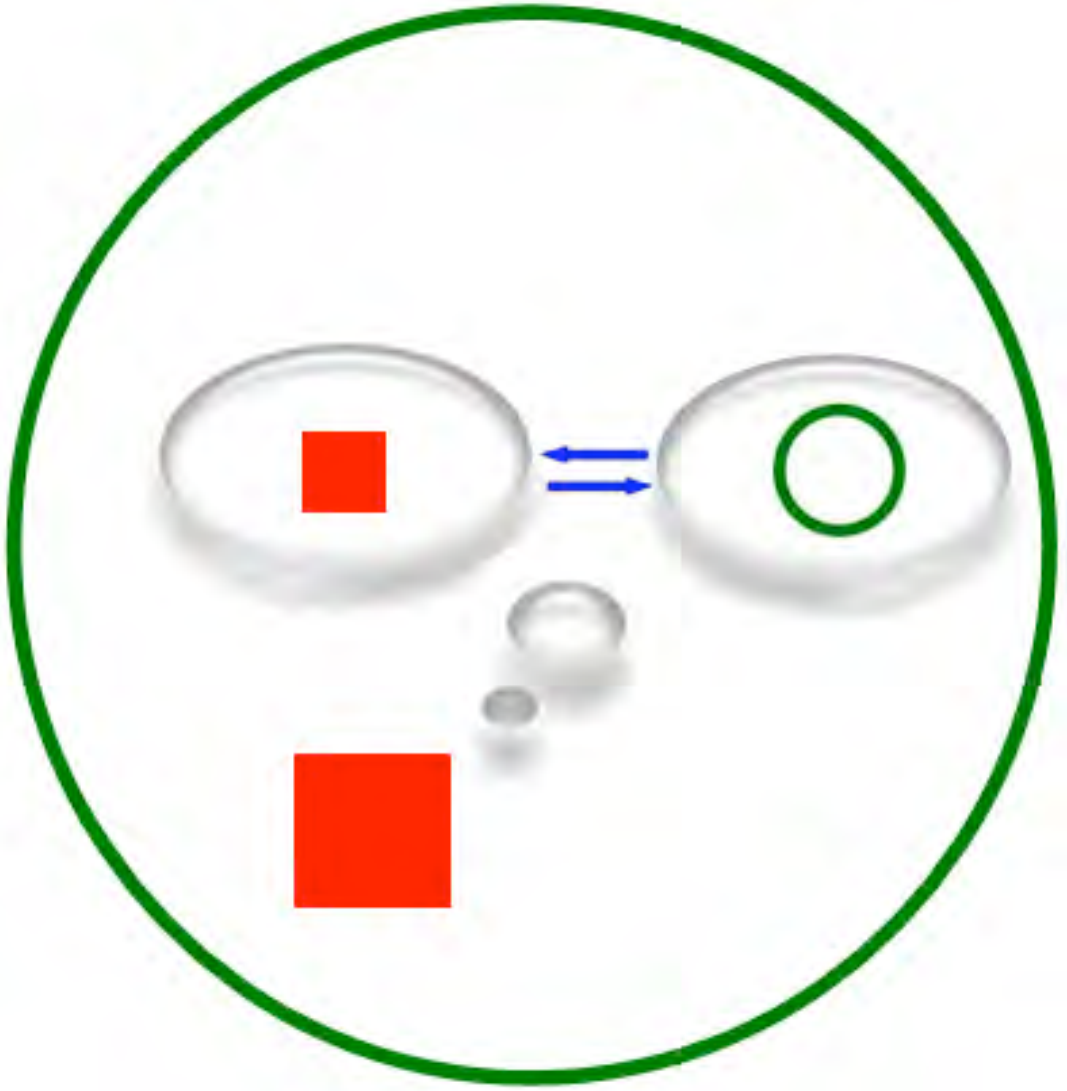
You attribute the body’s actions to your own agency (or not); this is ‘the illusion of conscious will’ (Daniel Wegner)



The ‘content’ of your consciousness is mostly secondary and illusory – it’s largely the consequences of keeping the planning system up to date, propagating knowledge through it, and evaluating current and future situations. You occasionally plan, ***but you can never act***.



And perhaps the cognitive peculiarities of consciousness are simply the natural operating properties of a system like this?



The world is not enough!

Many theories of consciousness concentrate on how well we can model the world.

I don’t believe that an internal model of the world alone is sufficient for consciousness. It will certainly give advanced cognition, but cognition is not consciousness for those of us who believe in the ‘hard problem’.

Navigation is not enough – it is necessary to act to change the world.

And if you don’t have a body worth modelling (e.g. if you are a Khepera robot) then you won’t develop an internal agent model, and so you won’t become conscious…

A proposal

One way to study these phenomena is to build a suitably complex robot, to embed it in a suitably complex environment and to examine the robot's behaviour and internal processes as it learns, evolves, or is designed to cope with its mission.

And to make sure any internal agent model developed is like our own, we should copy ourselves as best we can – our bodies, as well as our brains.

So how closely should we copy the body?

So how closely should we copy the body?

At least sufficiently closely to make it necessary to use motor programs (including those controlling eye movements) qualitatively similar to those used in the human body

So how closely should we copy the body?

At least sufficiently closely to make it necessary to use motor programs (including those controlling eye movements) qualitatively similar to those used in the human body

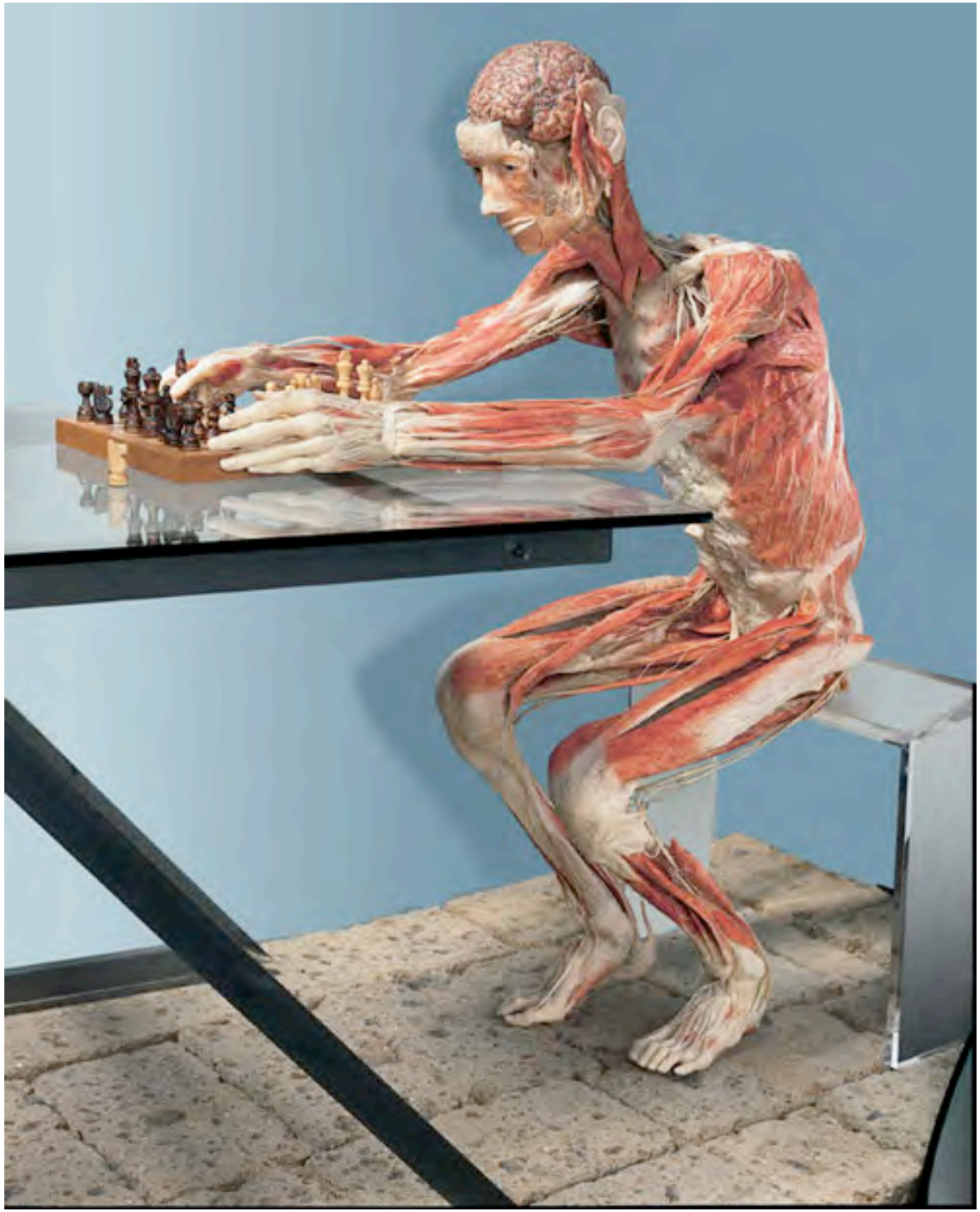
So what are bodies really like?

Not like this!

We’re familiar enough with the skeleton…

But most people have only become aware of what lies between the skin and the skeleton through the work of Gunther von Hagens.

‘Chess player’

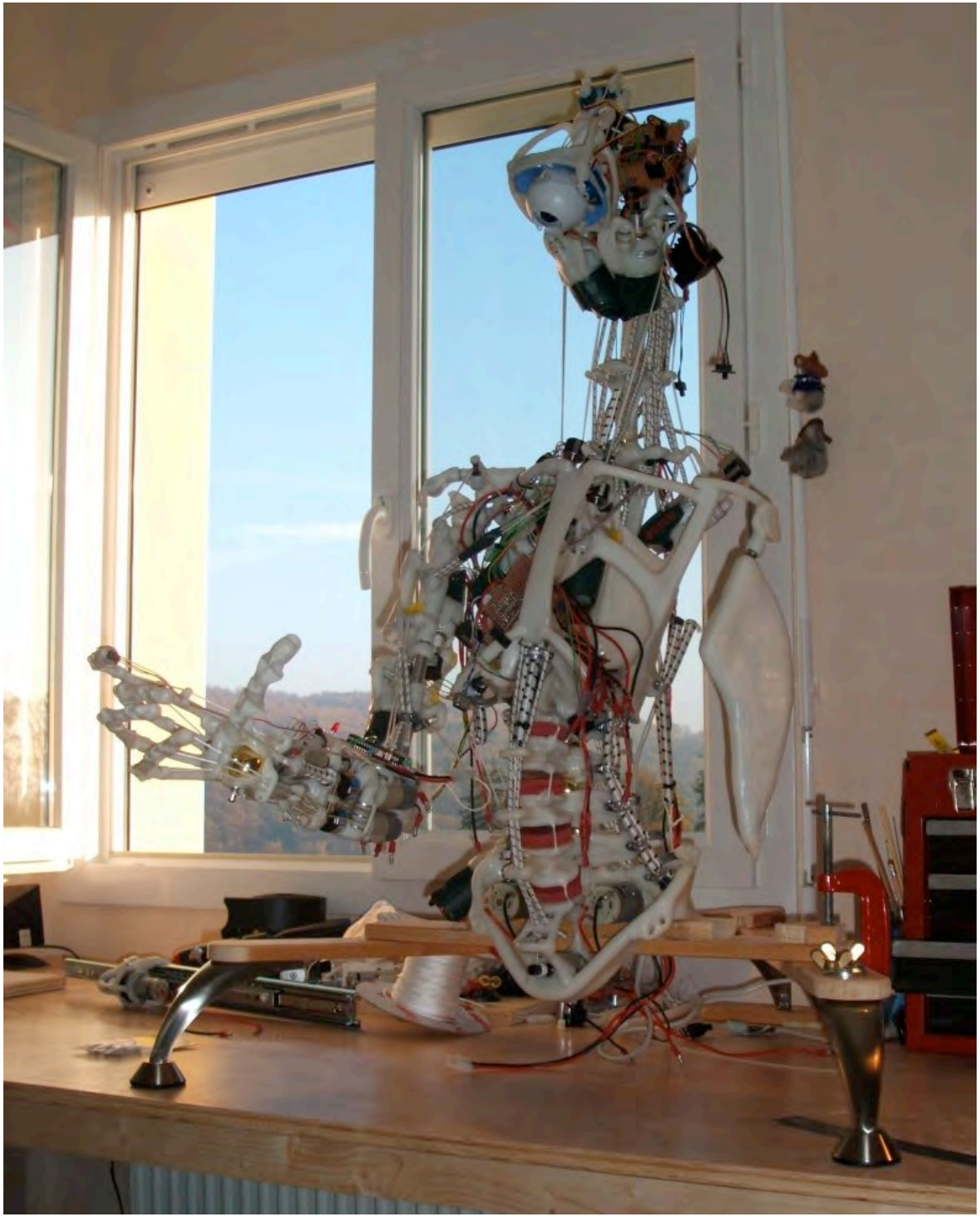


So how closely should we copy the body?

At least sufficiently closely to make it necessary to use motor programs (including those controlling eye movements) qualitatively similar to those used in the human body

And that means using **paired elastic actuators**, acting on a body consisting of **rigid elements** (bones) joined by **freely moving joints**, and linked by **passive elastic elements**…

…and you only have to start building robots like that to realise how different they are from ‘normal’ robots.



Dem bones, dem bones…

With these ***anthropomimetic*** robots, every movement and every external force is reflected through the whole structure, and they will deform the structure unless active compensation is applied

Dem bones, dem bones…

With these ***anthropomimetic*** robots, every movement and every external force is reflected through the whole structure, and they will deform the structure unless active compensation is applied.

Some of this compensation can be reactive, but much of it will have to be ***predictive*** (internal models again!) to enable actions to be carried out from a reasonably stable platform.

This goes far beyond merely maintaining the balance of a passively rigid structure.

MOVIES

Copying the brain as well

We’re also copying parts of the brain – those involved in early vision and the control of eye movements (work being done by Tom Troscianko, Iain Gilchrist, and Ben Vincent, Department of Psychology, University of Bristol)

How will we know if it's conscious?

I don't know, and shouldn't say. But other people are beginning to devise some useful frameworks for answering the question.

Igor Aleksander has proposed 5 axioms to define or characterise consciousness

Thomas Metzinger has identified 11 constraints on "…what makes a neural representation a phenomenal representation"

(T Metzinger, 2003: Being No-one: the self-model theory of subjectivity.)

Metzinger's 11 constraints

(1) Global availability
(2) Activation within a window of presence
(3) Integration into a coherent global state
(4) Convolved holism
(5) Dynamicity
(6) Perspectivalness
(7) Transparency
(8) Offline activation
(9) Representation of intensities
(10) "Ultrasmoothness": Homogeneity of simple content
(11) Adaptivity

We are aiming for a minimal notion of consciousness:

Metzinger: constraints 2, 3 and 7

(2) Activation within a window of presence

(3) Integration into a coherent global state

(7) Transparency

Aleksander's axioms

Damasio's core consciousness

What I left out

Architectures

Computational substrate

Social interaction

Language

Creativity

What I left out

Architectures (Grush, imagination, action selection)

Computational substrate (Spiking neural networks?)

Social interaction (No)

Language (No)

Creativity (?)

A warning from Thomas Metzinger

"Suffering starts on the level of Phenomenal Self Models. You cannot consciously suffer without having a globally available self-model. The PSM is the decisive neurocomputational instrument not only in developing a host of new cognitive and social skills but also in forcing any strongly conscious system to functionally and representationally appropriate its own disintegration, its own failures and internal conflicts…”

A warning from Thomas Metzinger

“Evolution is not only marvellously efficient but also ruthless and cruel to the individual organism. Pain and any other nonphysical kind of suffering, generally any representational state characterized by a "negative valence" and integrated into the PSM are now phenomenally owned. Now it inevitably, and transparently, is my own suffering. The melodrama, but also the potential tragedy of the ego both start on the level of transparent self-modeling. **Therefore, we should ban all attempts to create (or even risk the creation of) artificial and postbiotic PSMs from serious academic research."**

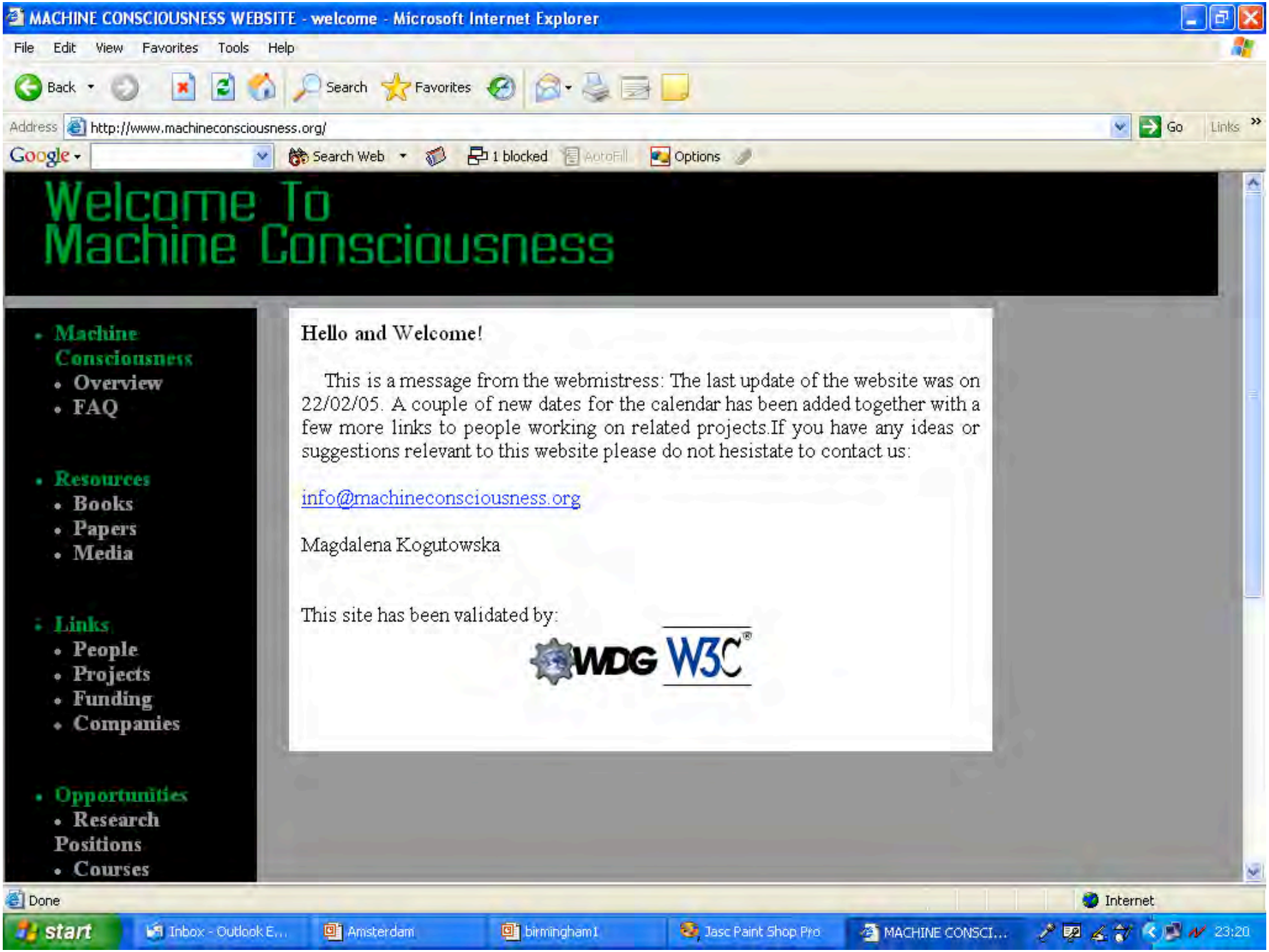
T. Metzinger, Being No-One (p 622).

For more information see

www.machineconsciousness.org

And please feel free to email me

owen@essex.ac.uk


MACHINE CONSCIOUSNESS WEBSITE - welcome - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.machineconsciousness.org/
Go Links
Google Search Web 1 blocked AutoFill Options
Welcome To
Machine Consciousness
Machine Consciousness
Overview
FAQ
Resources
Books
Papers
Media
Links
People
Projects
Funding
Companies
Opportunities
Research Positions
Courses
Hello and Welcome!
This is a message from the webmistress: The last update of the website was on 22/02/05. A couple of new dates for the calendar has been added together with a few more links to people working on related projects.If you have any ideas or suggestions relevant to this website please do not hesistate to contact us:
info@machineconsciousness.org
Magdalena Kogutowska
This site has been validated by:
WDG W3C
Done
Internet
start
Inbox - Outlook E...
Amsterdam
birmingham1
Jasc Paint Shop Pro
MACHINE CONSCI...
23:20