

Machine Consciousness and Creativity

Owen Holland



University of Essex

Department of
Computer Science

Roadmap

What is machine consciousness?

Where did consciousness come from?

The role of simulation

A robot-based approach

Evaluation and feeling

Why do the arts exist?

What is machine consciousness?

What is machine consciousness?

A new area of research dedicated to the construction of machines that are conscious like you...

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...

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

It's not about creating the illusion that machines are conscious – that is depressingly easy

Can you imagine consciousness without feeling?

Can you imagine consciousness without feeling?

Could Spock exist?

Can you imagine consciousness without feeling?

Could Spock exist?

Does a consciousness without feeling seem alien to us?

Can you imagine consciousness without feeling?

Could Spock exist?

Does a consciousness without feeling seem alien to us?

Would machine consciousness without feeling be consciousness as we know it?

How could we build a conscious machine?

How could we build a conscious machine?

(1) Identify the components of consciousness, and implement all of them in a machine?

How could we build a conscious machine?

(1) Identify the components of consciousness, and implement all of them in a machine?

(2) Identify the components of the machine that produces consciousness (the brain) and copy them?

How could we build a conscious machine?

(1) Identify the components of consciousness, and implement all of them in a machine?

(2) Identify the components of the machine that produces consciousness (the brain) and copy them?

(3) Identify the circumstances in which consciousness arose, copy them, and hope that consciousness emerges again?

How could we build a conscious machine?

(1) Identify the components of consciousness, and implement all of them in a machine?

(2) Identify the components of the machine that produces consciousness (the brain) and copy them?

(3) ***Identify the circumstances in which consciousness arose, copy them, and hope that consciousness emerges again?***

How did consciousness arise?

How did consciousness arise?

We don't know (and of course 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 (and of course 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 (and of course 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

How did consciousness arise?

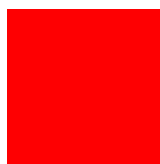
We don't know (and of course we don't know exactly what consciousness *is*) but it was probably something to do with the development of high intelligence

How did intelligence arise?

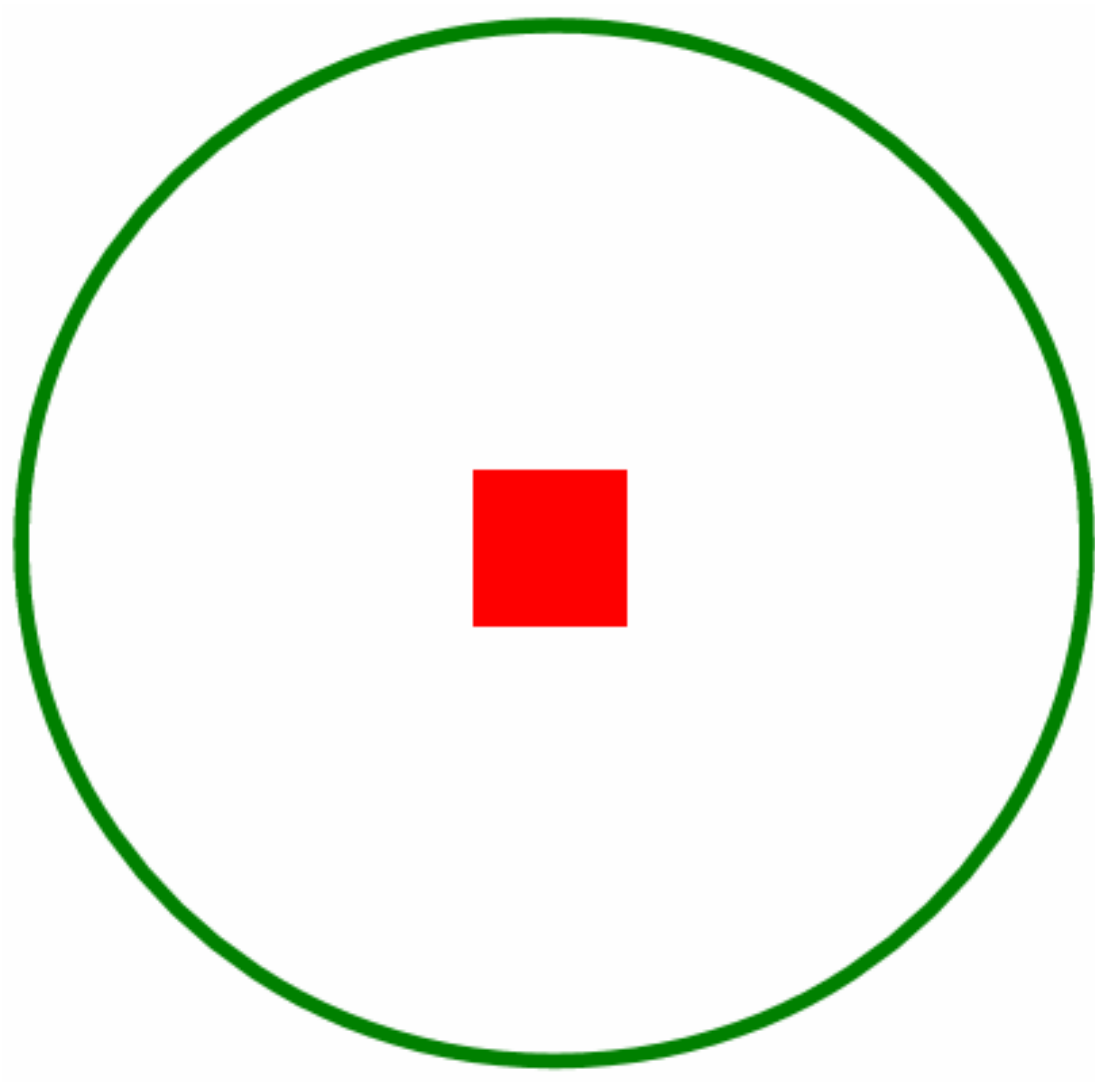
Through natural and ***sexual selection***

(If you haven't read it yet, have a look at Geoffrey Miller's 'The Mating Mind')

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 achieve its task (or mission)?

- by being preprogrammed for every possible contingency?

How could the agent achieve its task (or mission)?

- by being preprogrammed for every possible contingency? No

How could the agent achieve its task (or mission)?

- 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?

How could the agent achieve its task (or mission)?

- 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

How could the agent achieve its task (or mission)?

- 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?

How could the agent achieve its task (or mission)?

- 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

Otherwise, 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.

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.

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 working models 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.

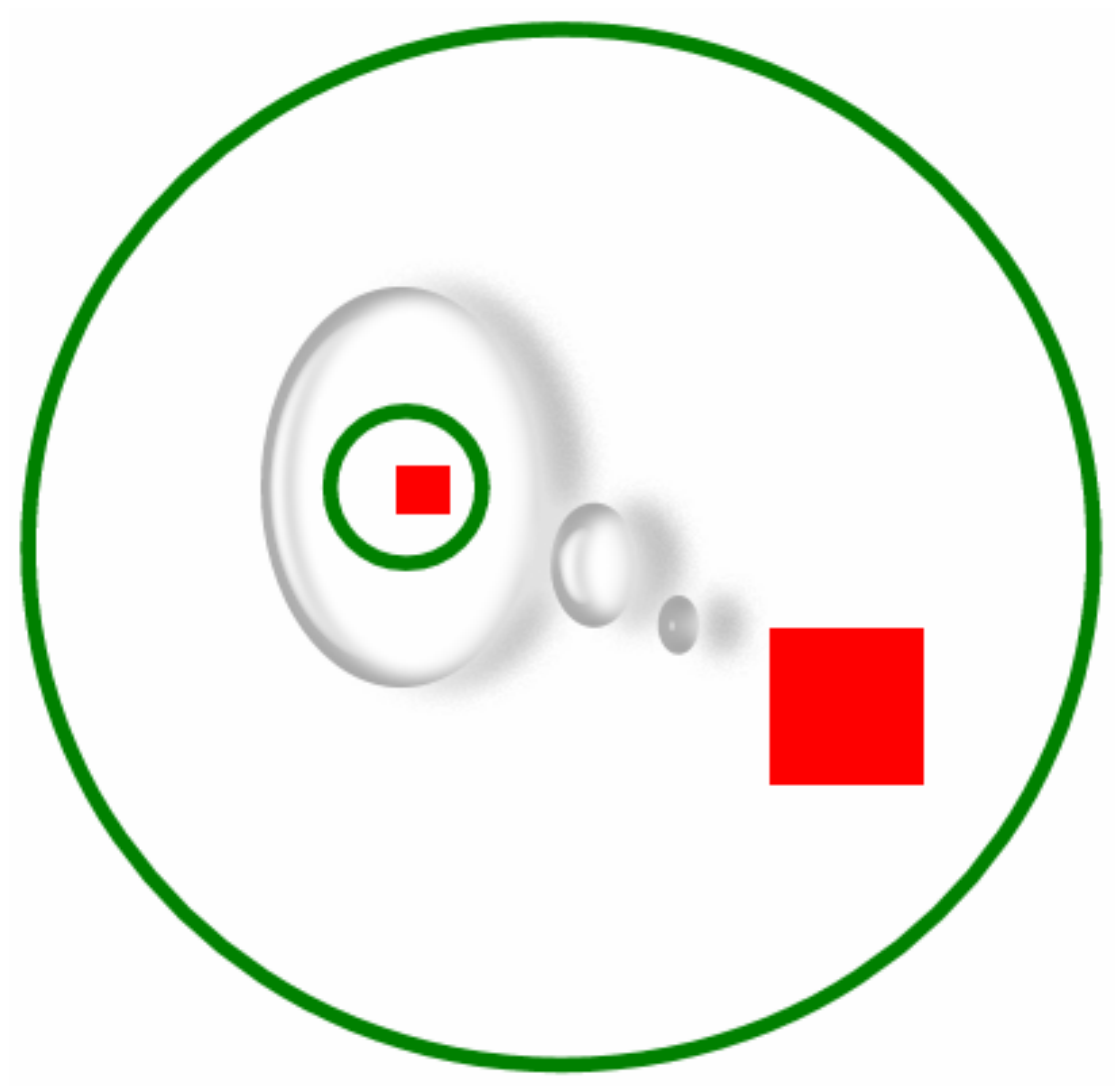
The **world** is different. It is 'complex, occasionally novel, dynamic, and hostile'. It's only locally available, and may contain objects of known and unknown kinds in known and unknown places.

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?

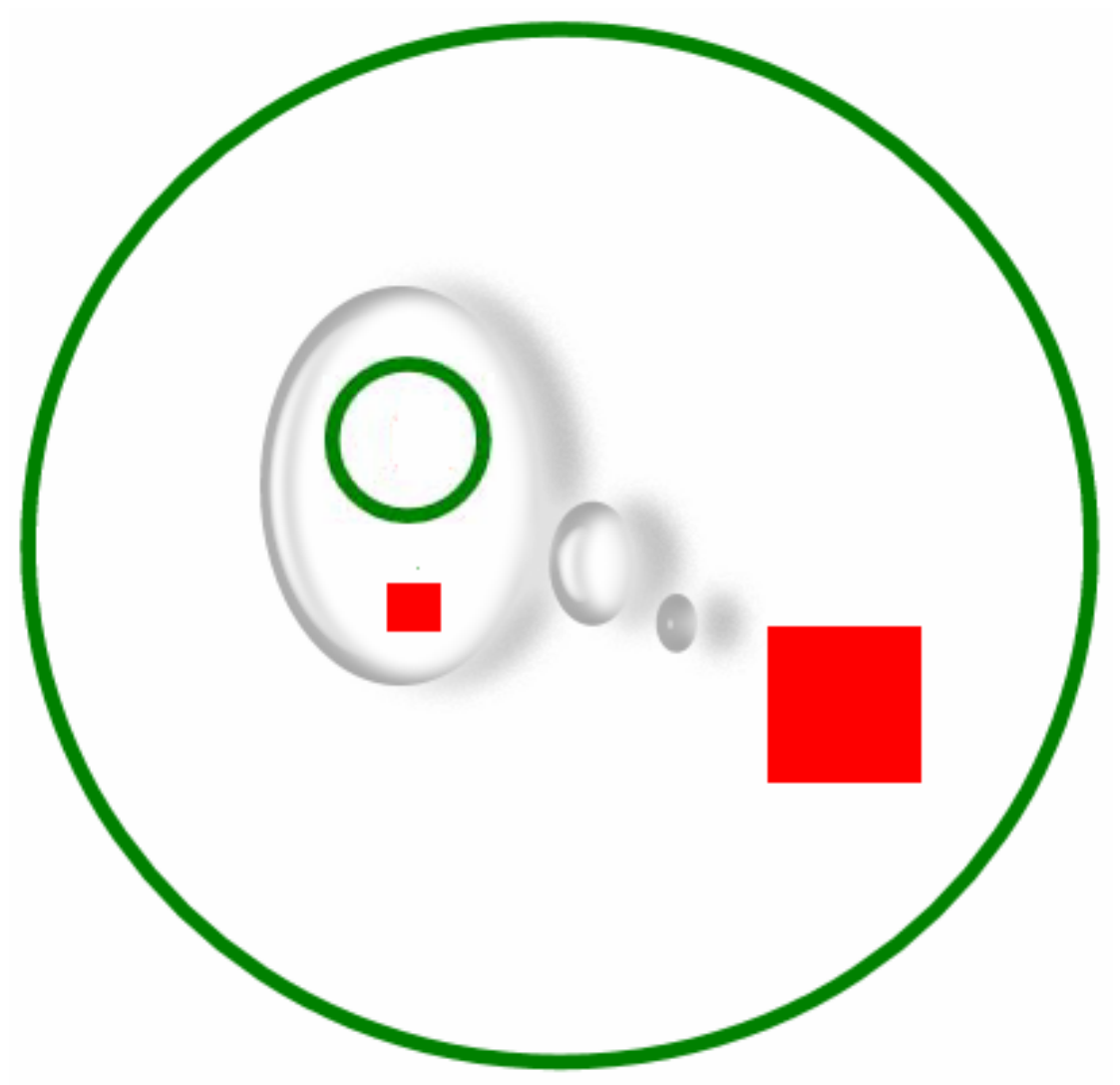


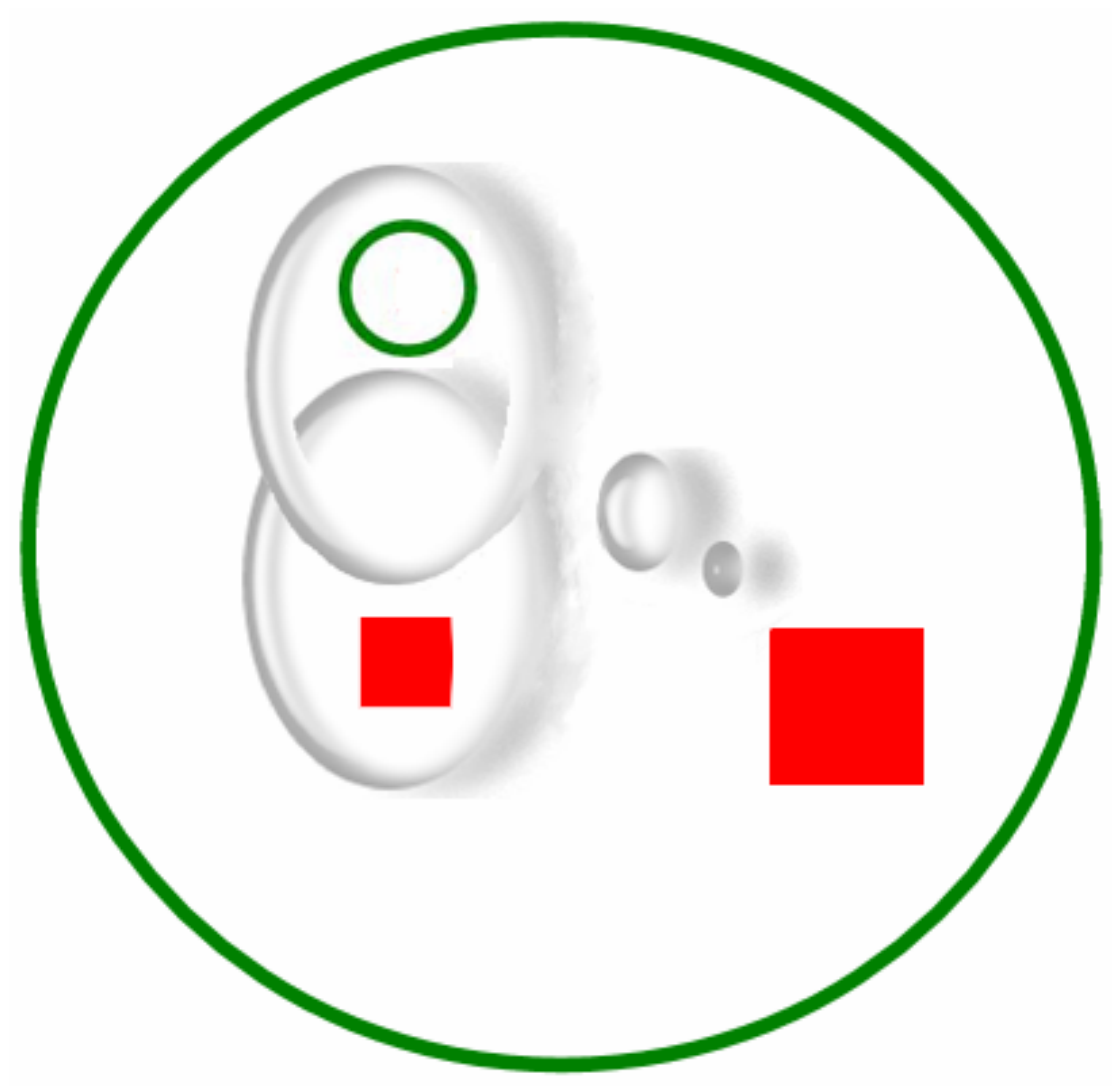
What is needed for simulation?

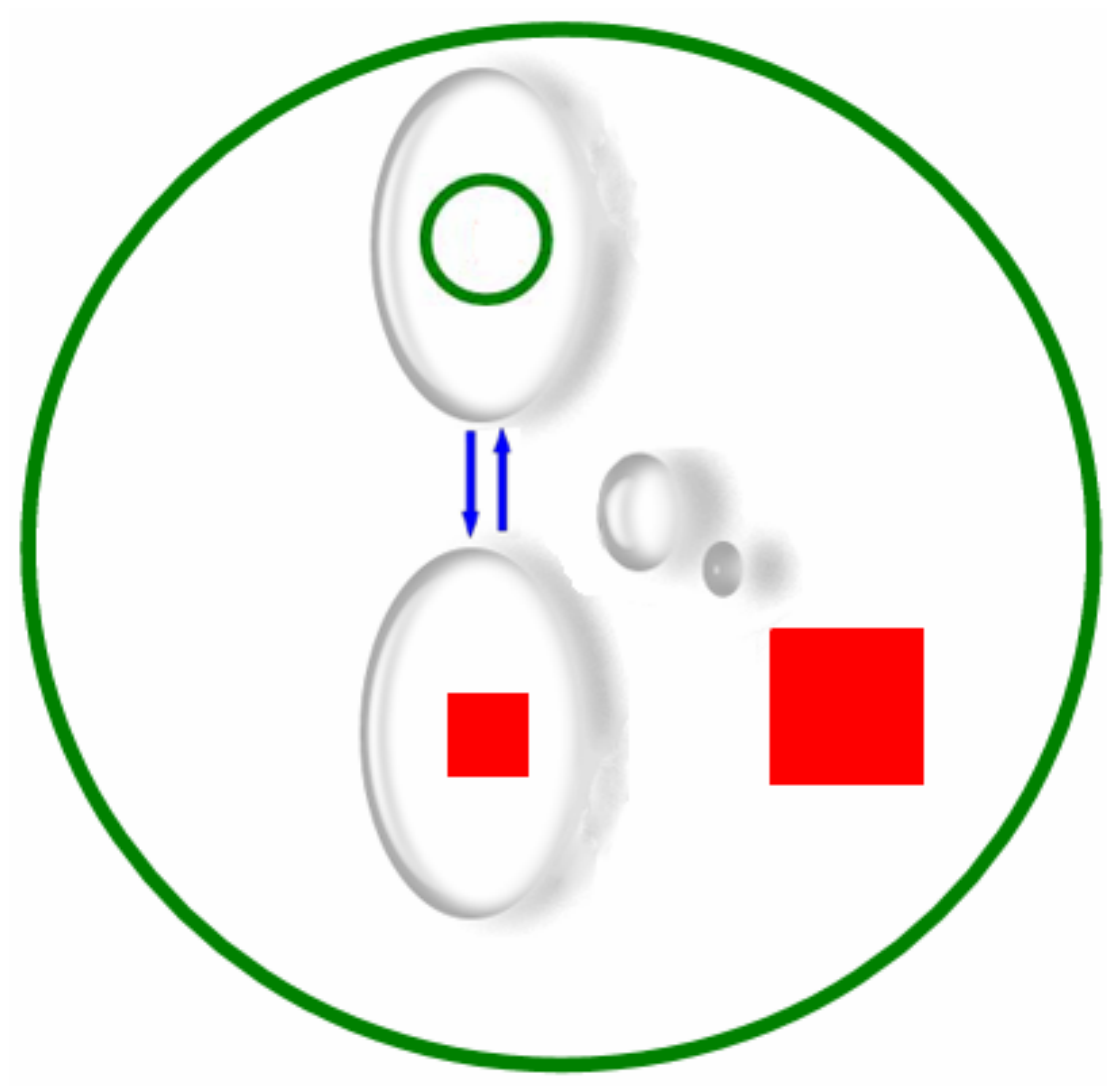
The **body** is always present and available, and changes slowly, if at all. When it moves, it is usually because it has been commanded to move.

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 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 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 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.***”

.

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.***”

How about ‘...a model of the whole machine, not just the brain’?

In other words...

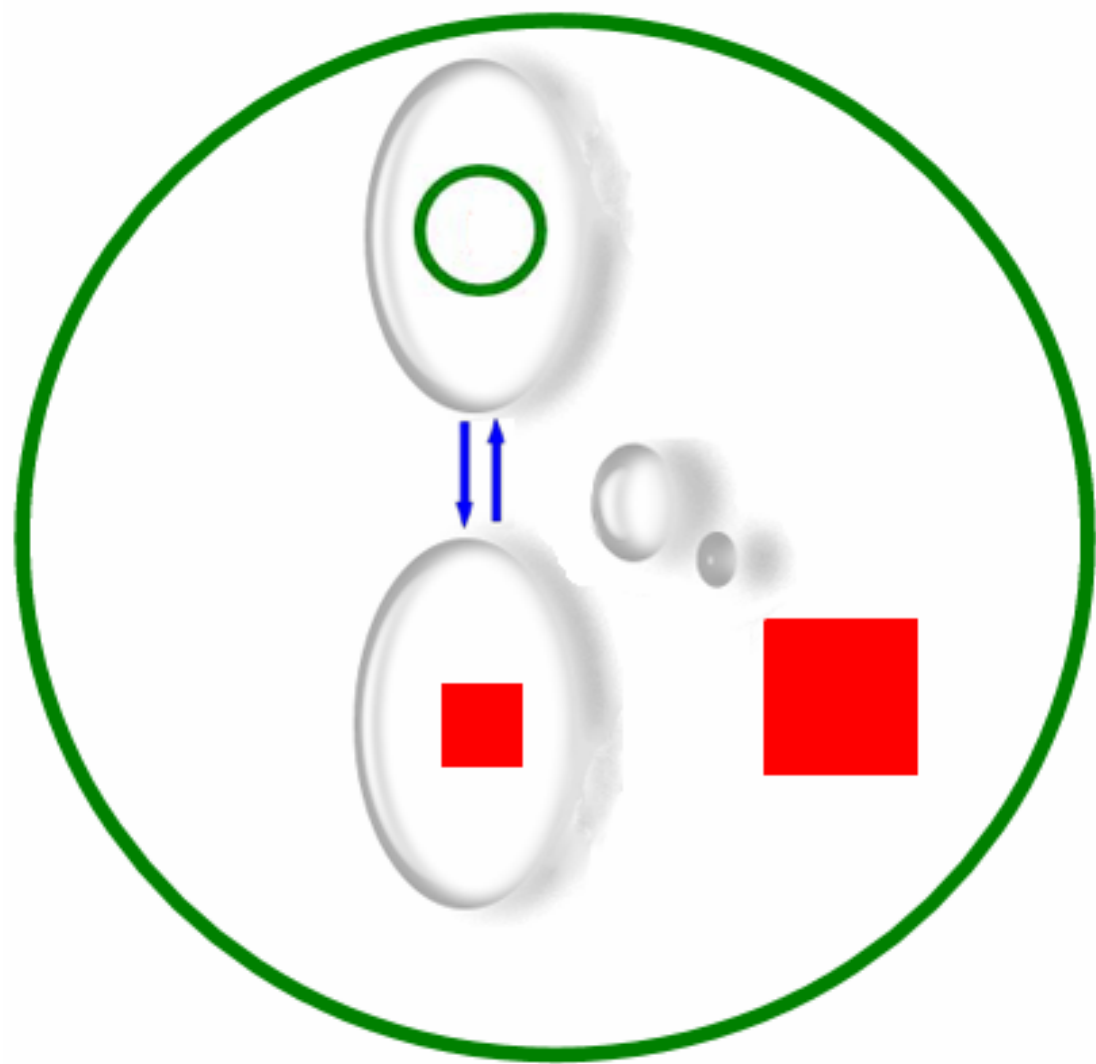
Intelligence may depend on the possession and manipulation of an internal model of the agent (the IAM) interacting with an internal model of the world

AND

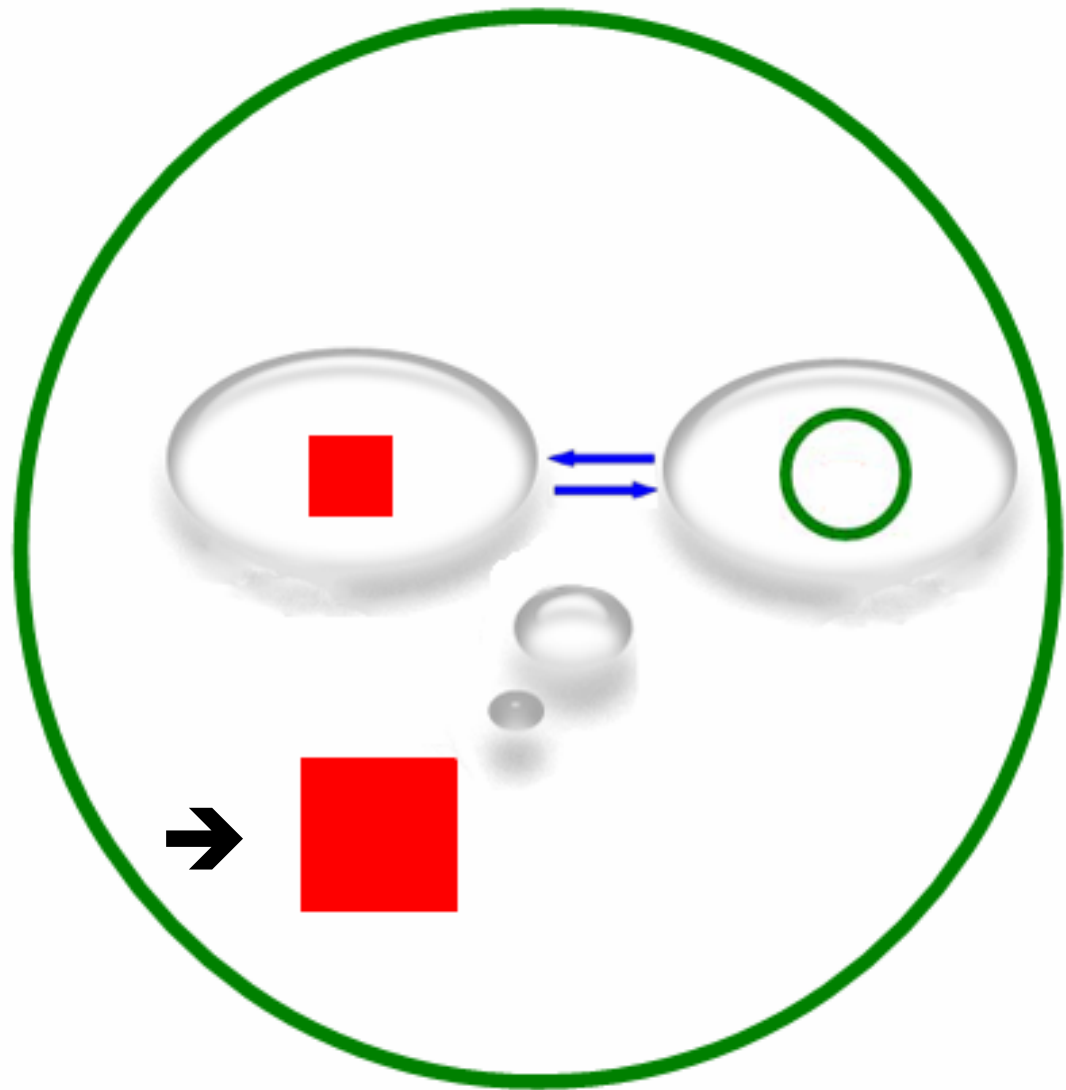
the presence and interaction of these models may also underlie the production of consciousness.

A hypothesis

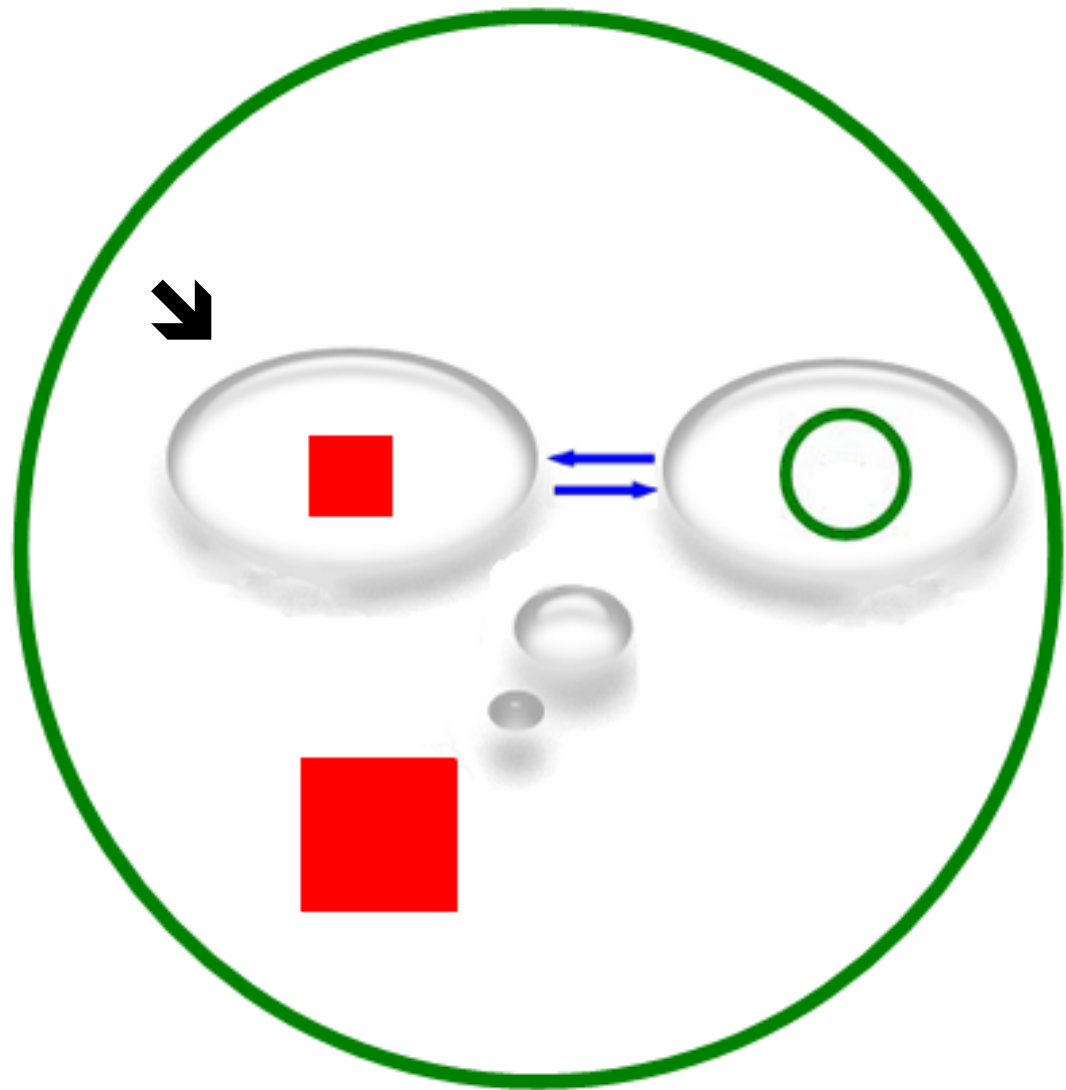
In humans (and some animals?) it is the IAM – the internal agent model – that is conscious, not the agent itself.



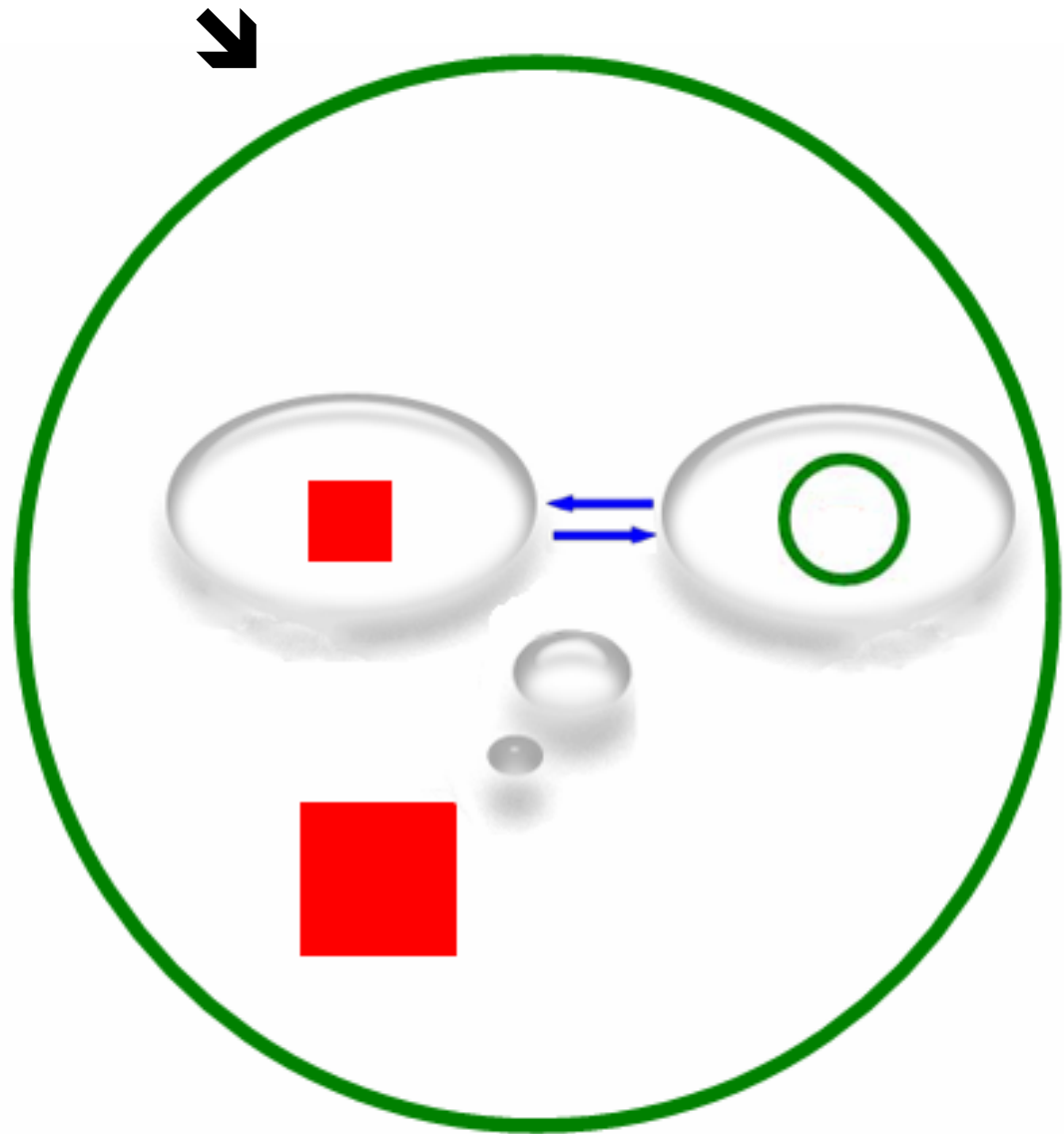
You think
you're this...



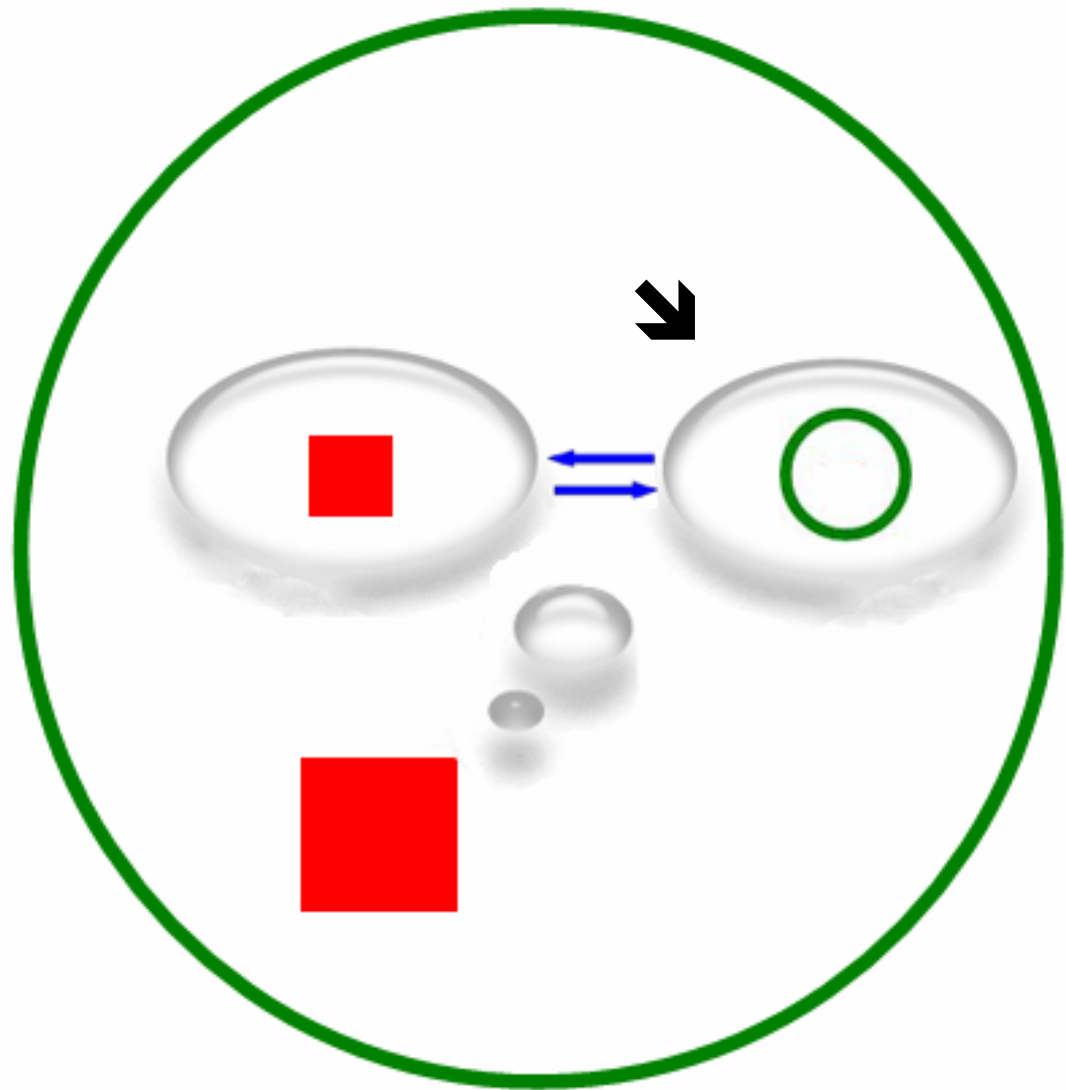
...but actually
you're this – a
model of yourself



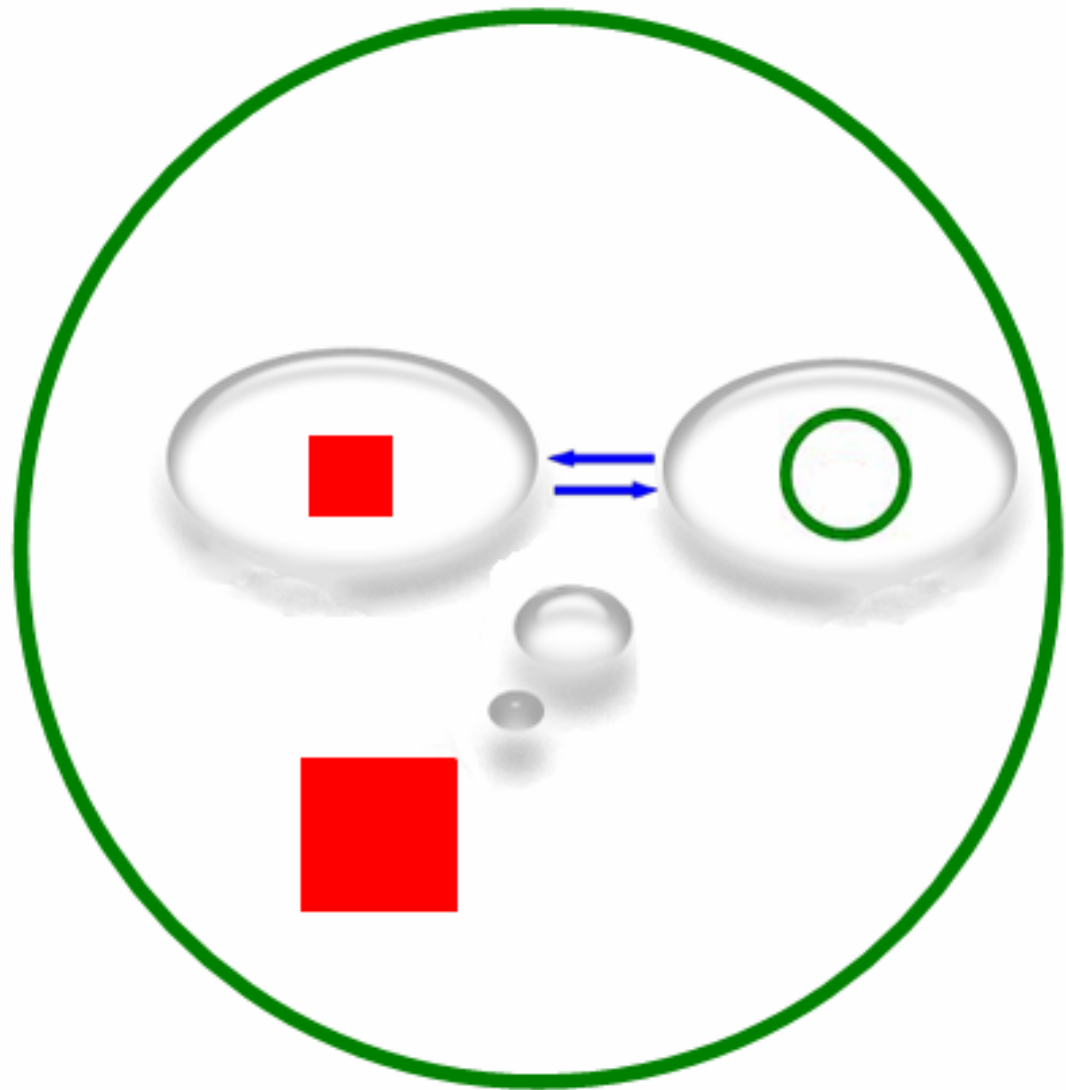
You think you
experience the
real world



But actually you
experience a
model of the real
world

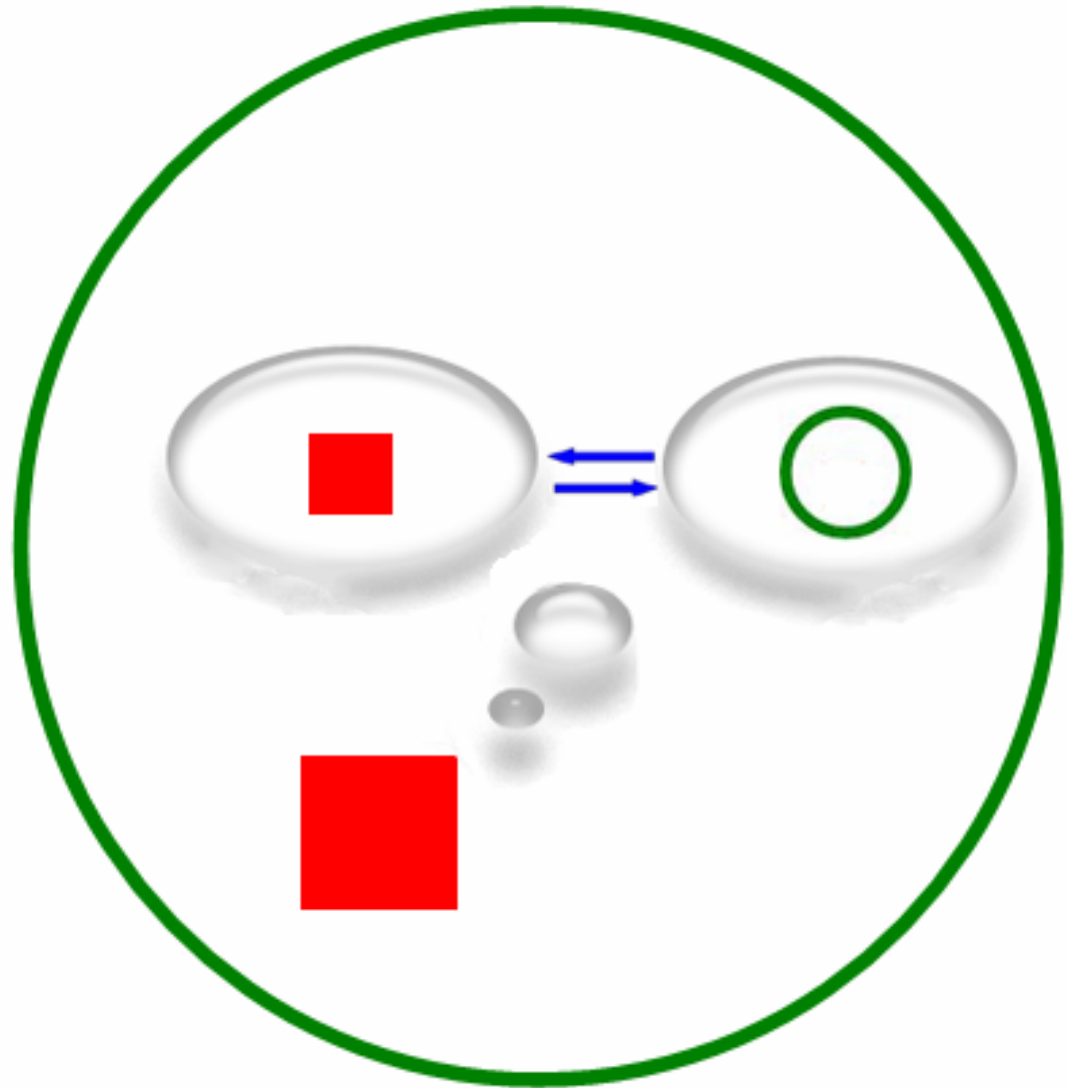


Consciousness and feelings are in the IAM – the system's 'software' model of itself...

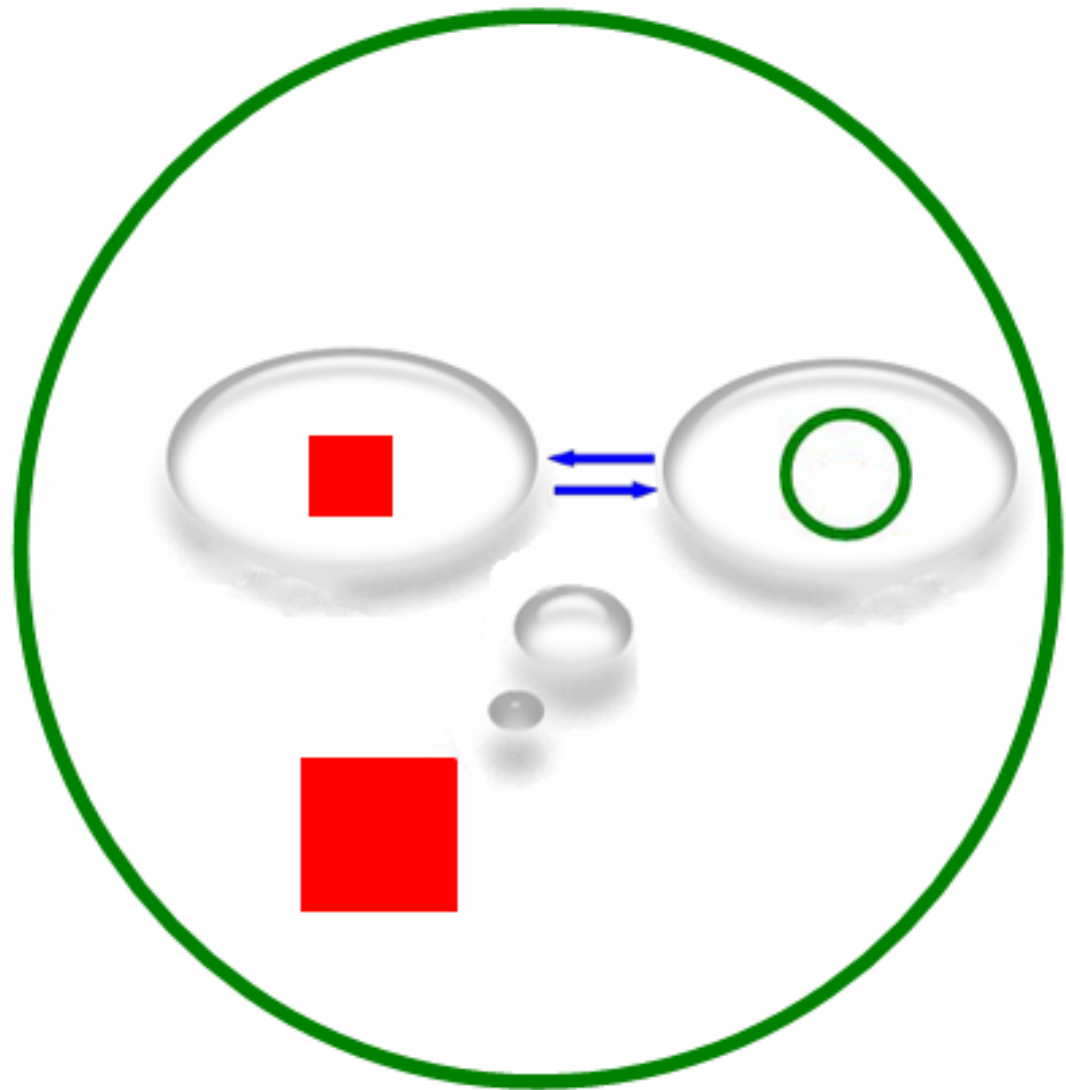


Consciousness and feelings are in the IAM – the system's 'software' model of itself...

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

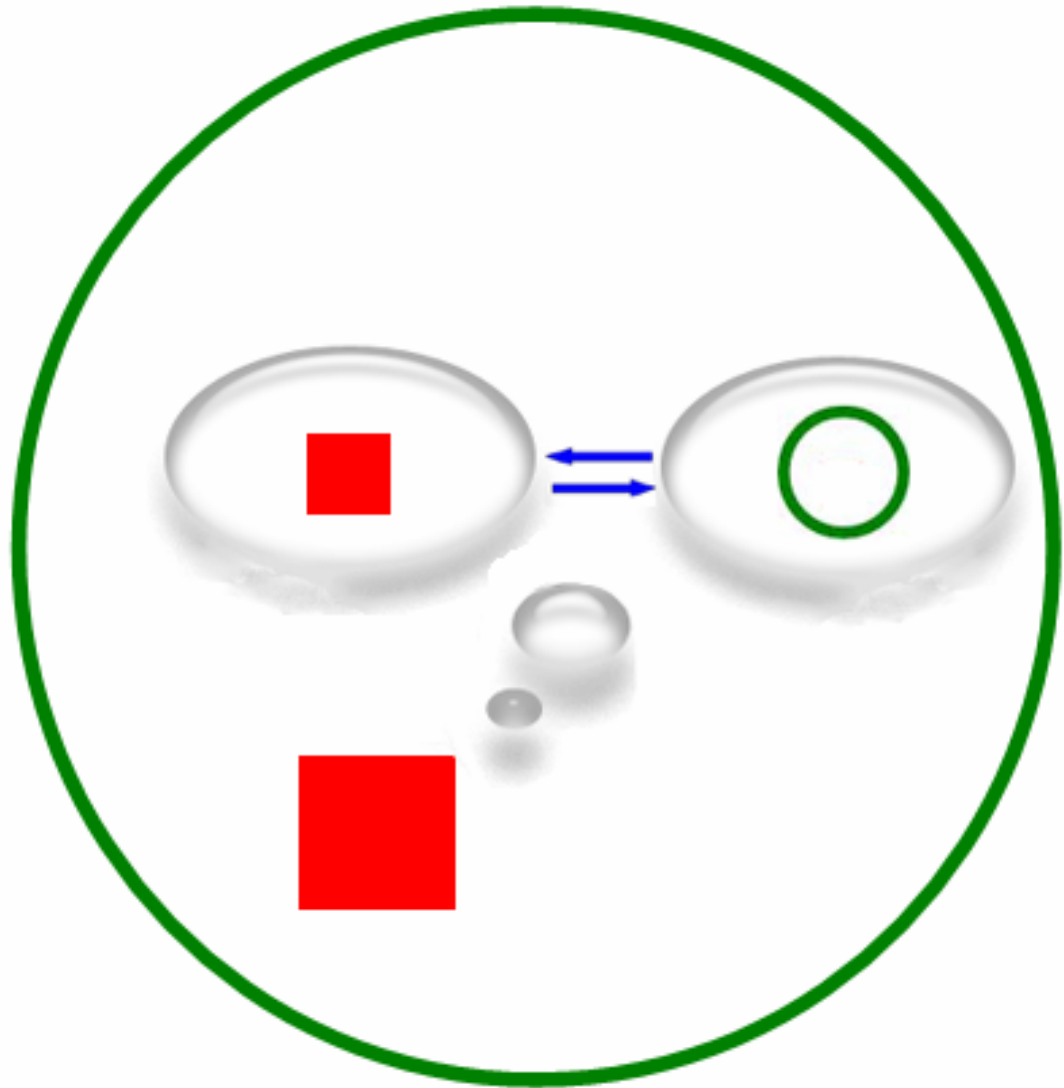


You think you
control your body,
and act on the real
world

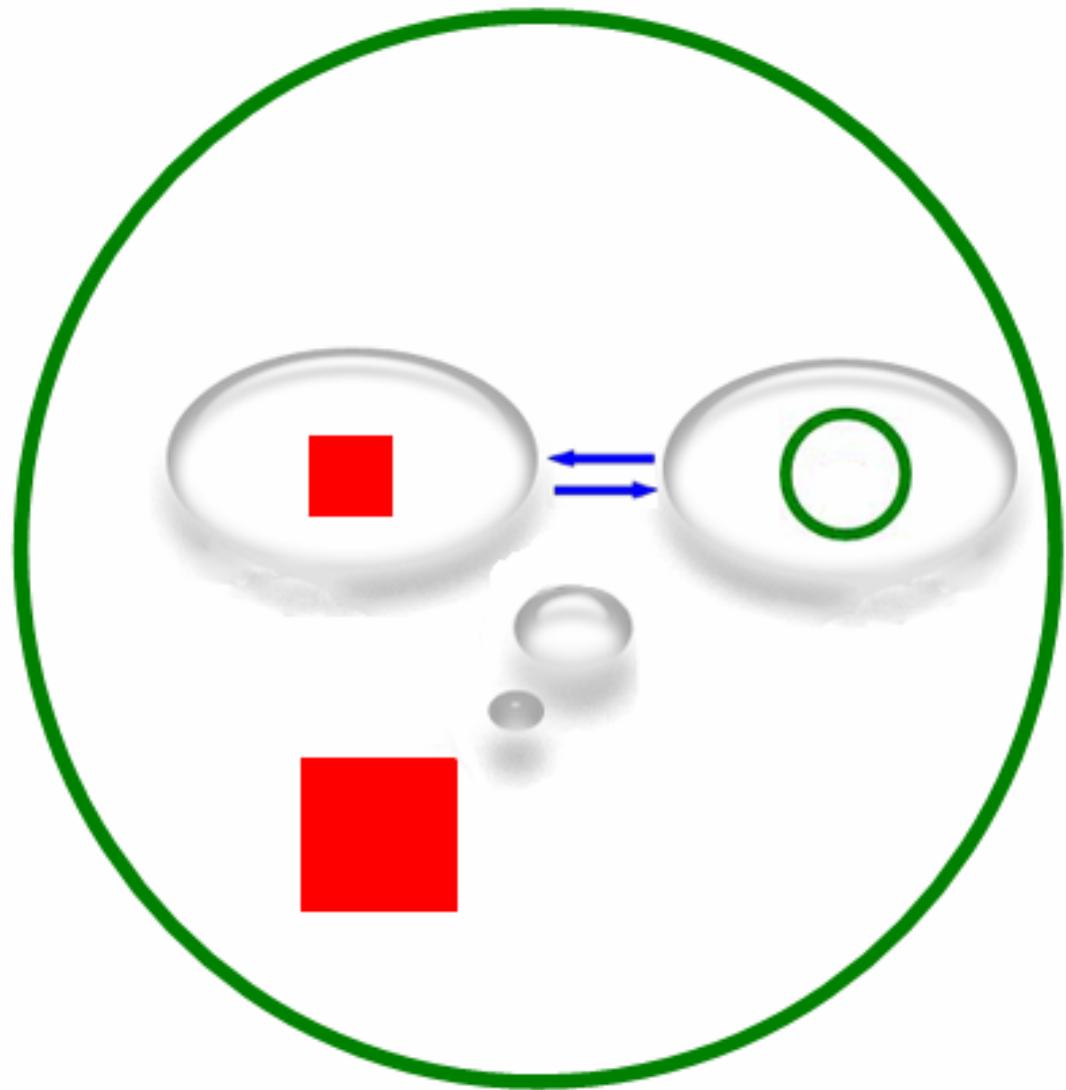


But your body is controlled by other structures within your brain, using the information about 'good' choices.

You attribute its 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, and propagating knowledge through it. You occasionally plan, but you can never act.



A proposal

The 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 to cope with its mission.

A proposal

The 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 to cope with its mission.

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

A proposal

The 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 to cope with its mission.

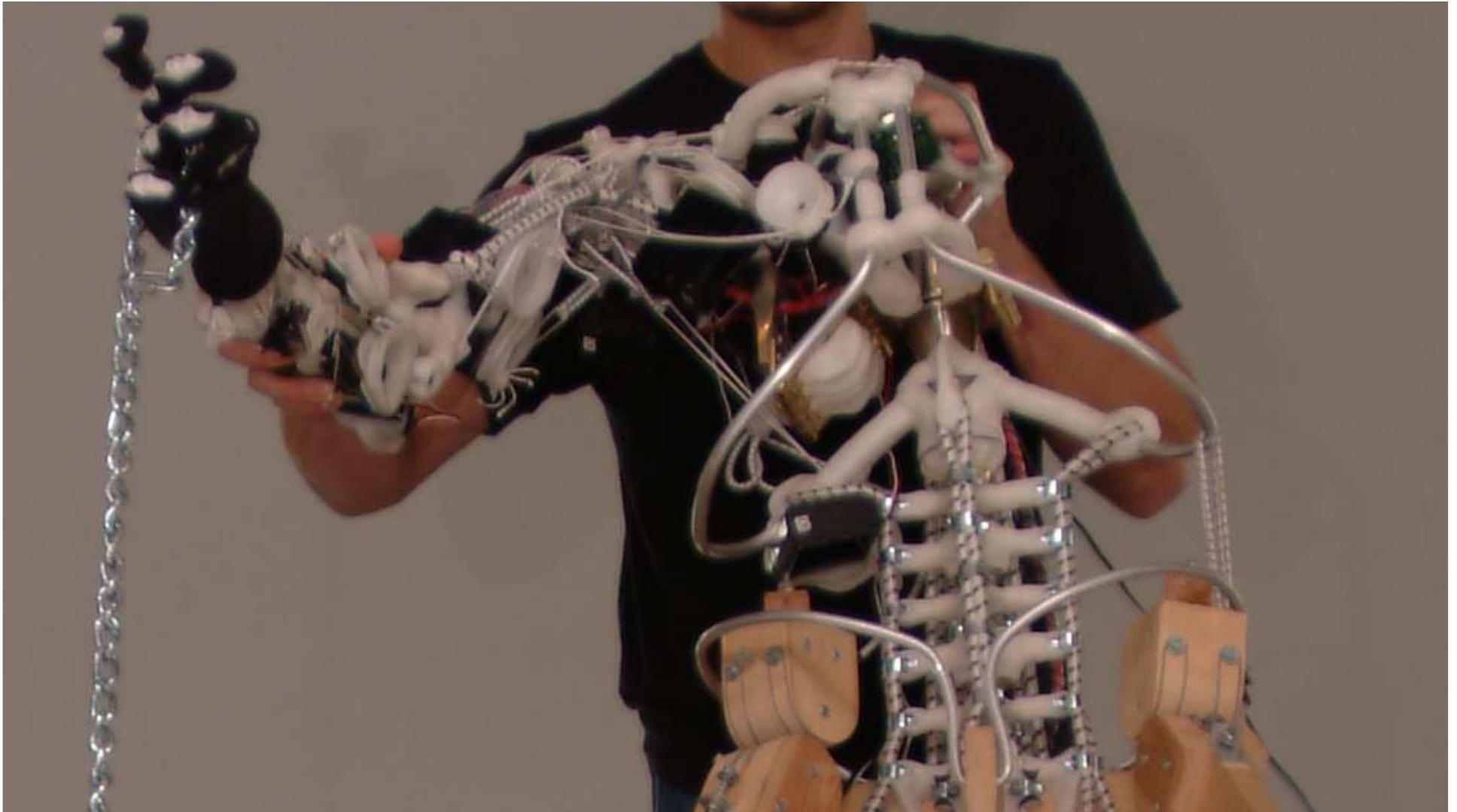
And to make sure any consciousness developed is like our own, we should copy ourselves as best we can – our bodies, as well as our brains.

An acceptance

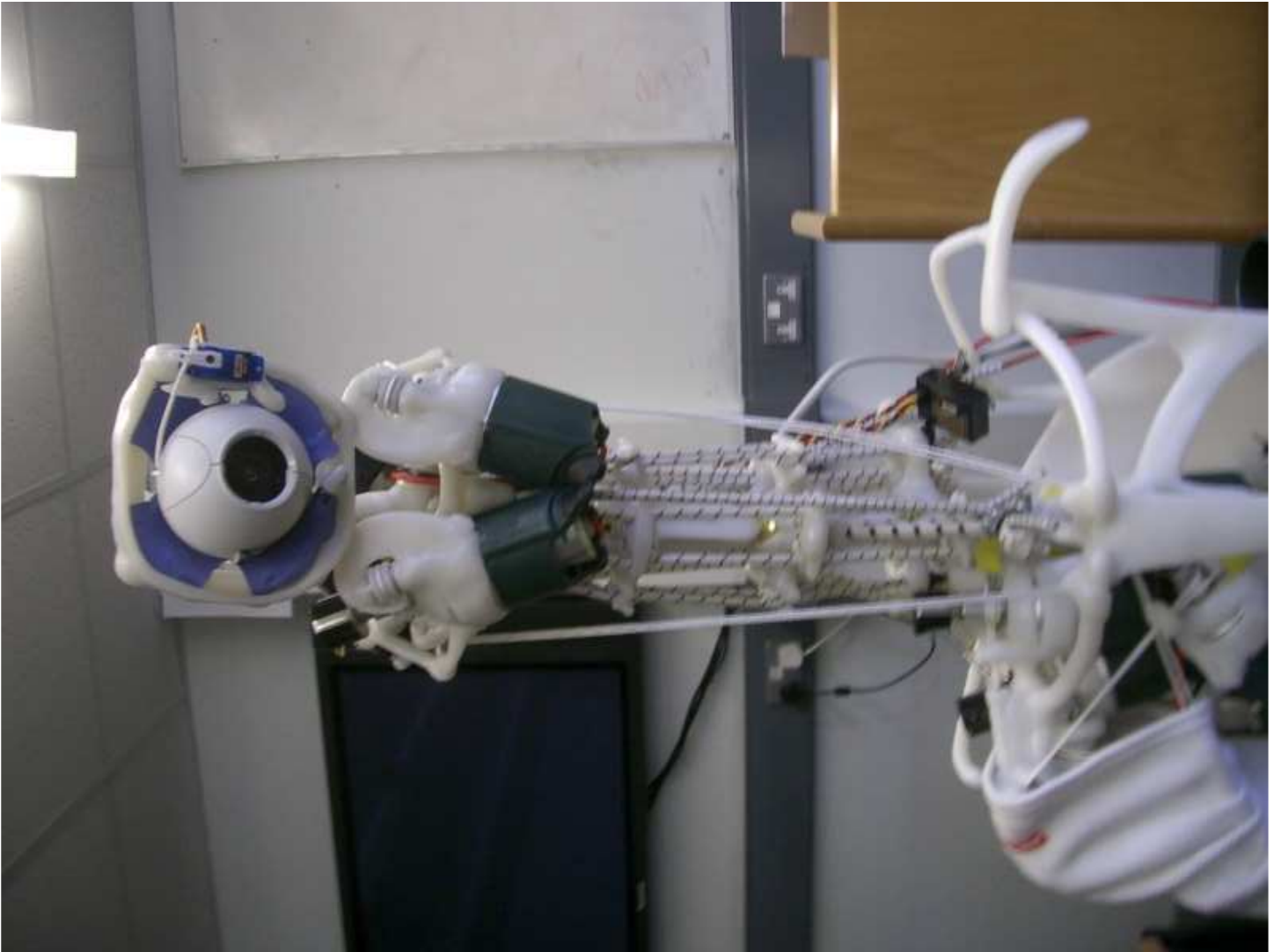
EPSRC Adventure Fund agreed!

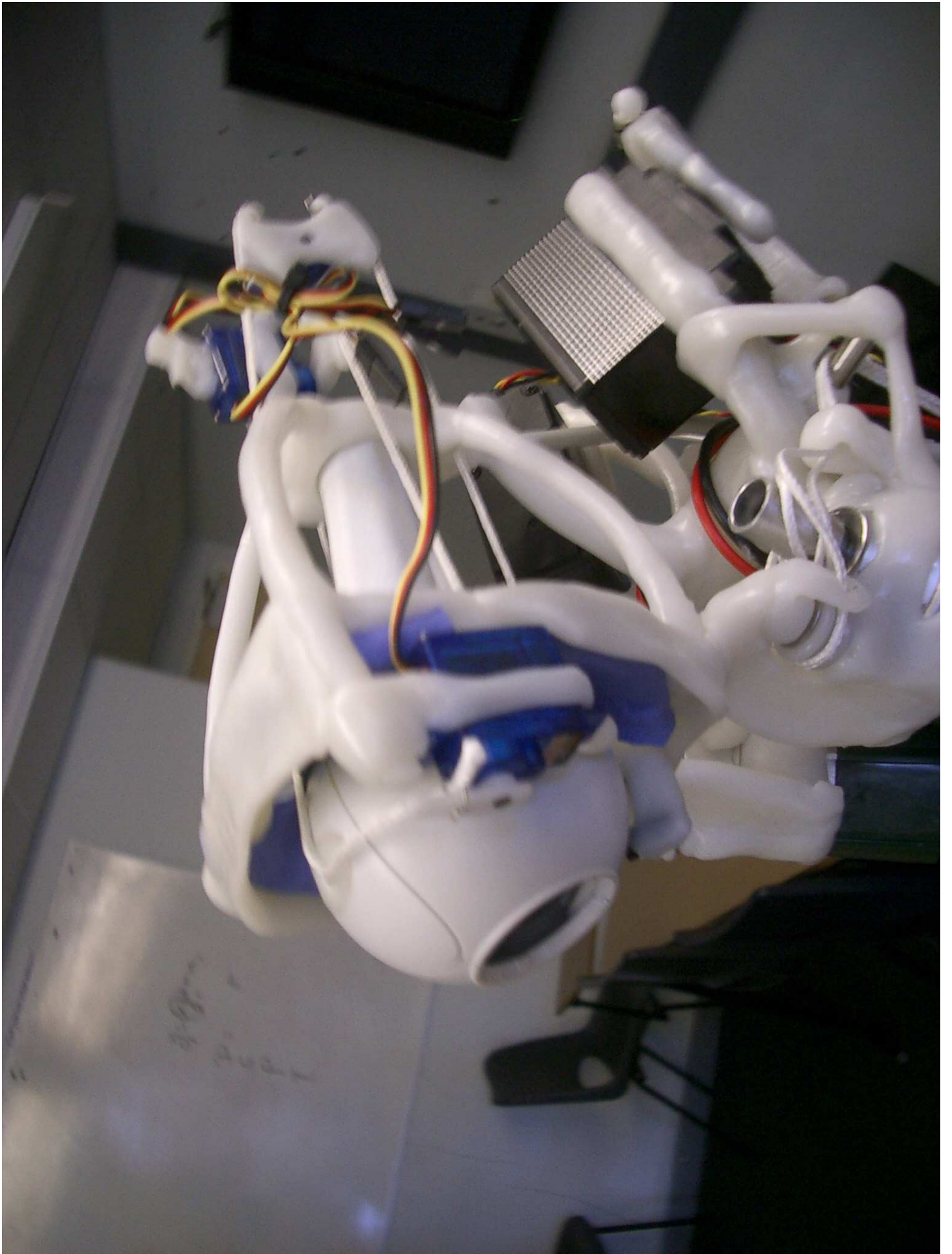
Copying the body





CRONOS 1





So how closely should we copy the body?

So how closely should we copy the body?

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?

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...

So how closely should we copy the body?

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 ***anthropomorphic*** 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 ***anthropomorphic*** 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.

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

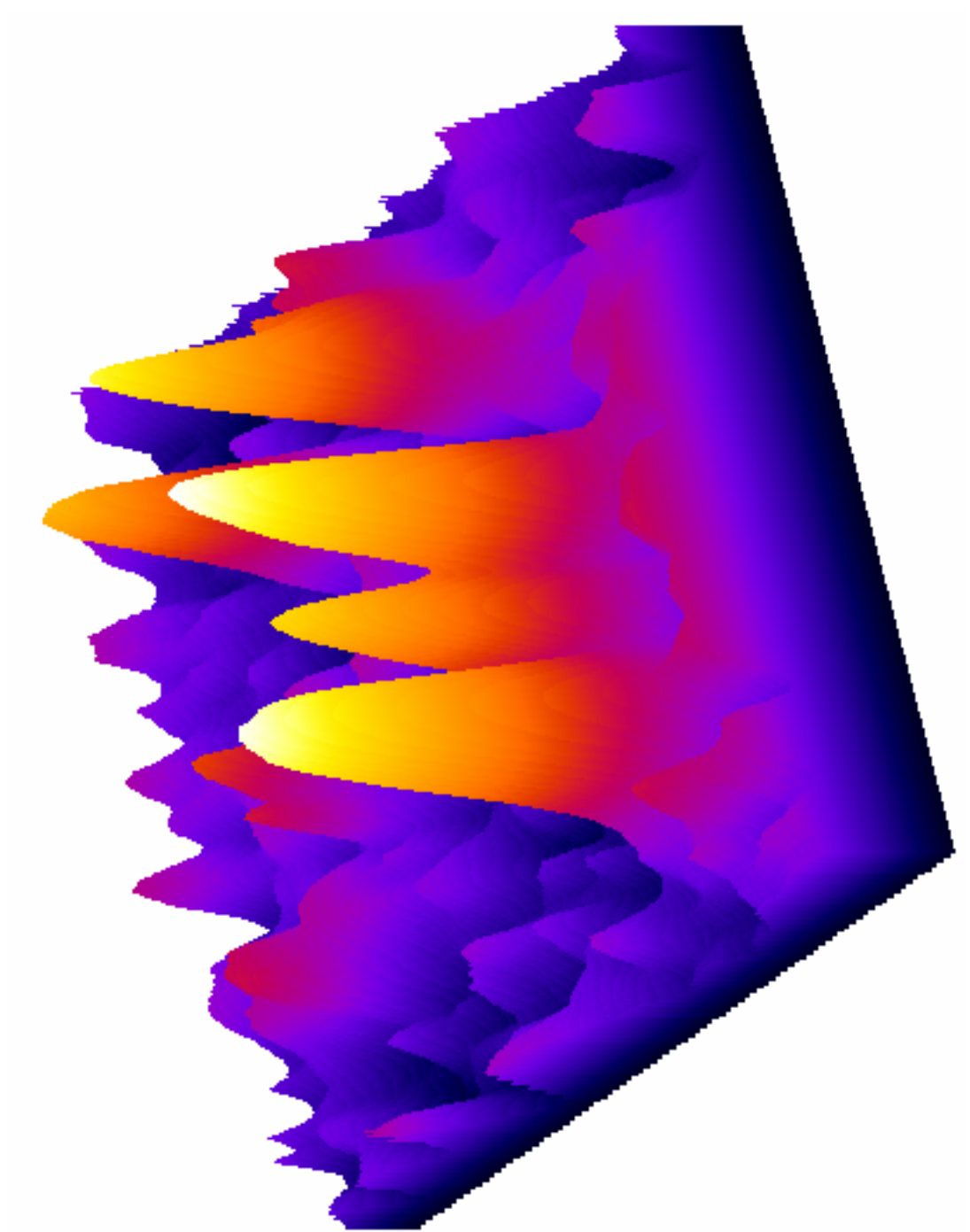
This goes far beyond merely maintaining the balance of a passively rigid structure.

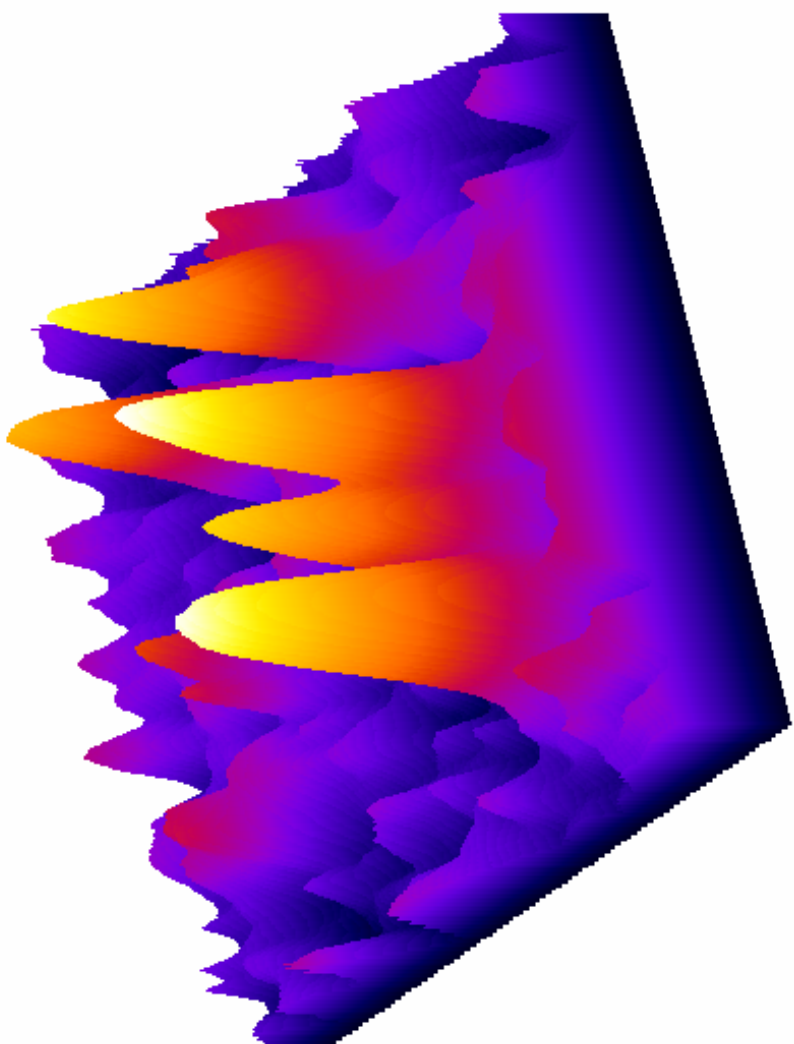
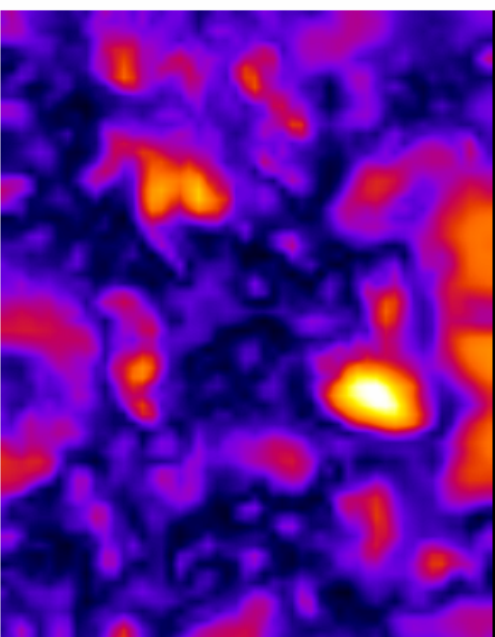
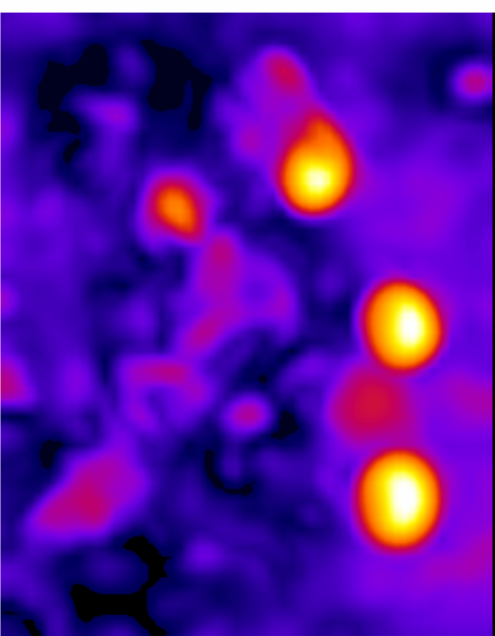
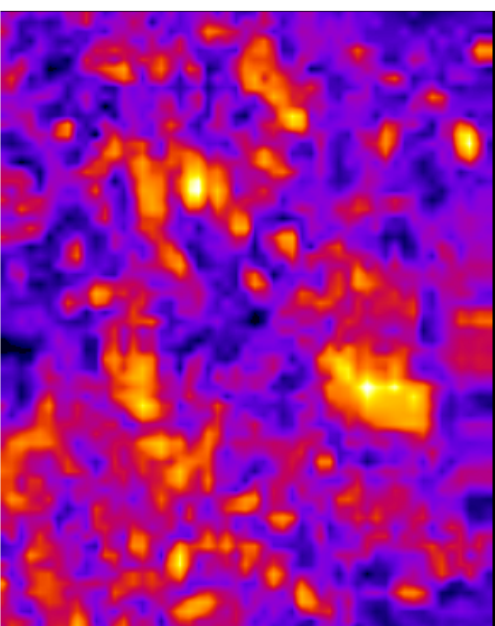
Copying the brain as well

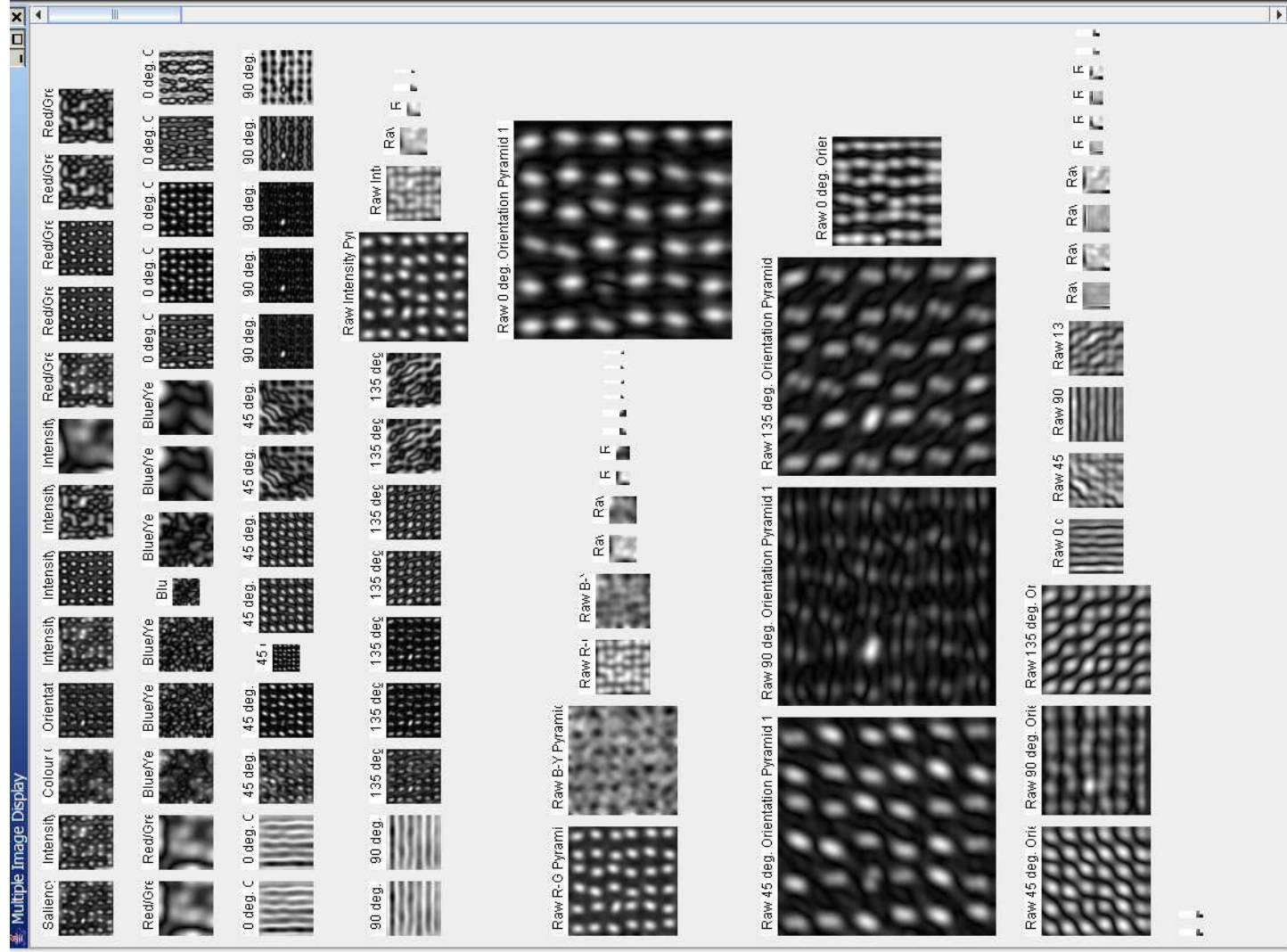
We're also copying parts of the brain – those involved in early vision and the control of eye movements (work being done by Tom Troscianko and Iain Gilchrist, Department of Psychology, University of Bristol)

To get started, we're using saliency mapping









Planning, imagination, and creativity

Formally, what we are building is a planning system. Previous attempts (GOF AI) used a symbolic substrate, and developed a huge range of methods – more or less mechanical – for generating the actions to be evaluated.

Planning, imagination, and creativity

Formally, what we are building is a planning system. Previous attempts (GOFAI) used a symbolic substrate, and developed a huge range of methods – more or less mechanical – for generating the actions to be evaluated.

We know **ALMOST NOTHING** about using a non-symbolic simulation engine as the substrate of a planning system (there is some work using noise to search within learned neural network representations) but we'll take as much as we can from GOFAI.

Evaluation and feeling

A key component of planning and imagination based systems is the evaluation function – the part of the architecture that assesses real and imagined situations for the contribution they will make to the mission.

Almost all researchers in the area – e.g. Rolls, Edelman, Aleksander, and ourselves – are agreed that this assessment is expressed in terms of a single quantity located on a continuum ranging from positive to negative.

We also suspect that, for any such system, the evaluation of an imagined or hypothetical situation or action is expressed in real terms – for example, it is genuinely unpleasant to imagine an unpleasant situation.

We also suspect that, for any such system, the evaluation of an imagined or hypothetical situation or action is expressed in real terms – for example, it is genuinely unpleasant to imagine an unpleasant situation.

We further suspect that the evaluation system runs constantly – that is, it will constantly evaluate whatever is offered to it, whether real, imagined, concrete, abstract, possible, impossible, current or remembered, and will express that evaluation in the same currency.

In a conscious machine, it is the relationship between the real or imaginary input to such a system and this evaluative output that will provide the substrate for feeling.

Like us, it cannot be neutral – every input, whether from the world, imagination, or memory, will give rise to an evaluative response, which we experience as feeling, and use to judge.

In a conscious machine, it is the relationship between the real or imaginary input to such a system and this evaluative output that will provide the substrate for feeling.

Like us, it cannot be neutral – every input, whether from the world, imagination, or memory, will give rise to an evaluative response, which we experience as feeling, and use to judge.

Can we say what will underpin these feelings?

Yes – relevance to the original mission

A puzzle for evolution: Why do the arts exist?

Within the context of evolutionary theory, the arts have no obvious natural selection value...

A puzzle for evolution: Why do the arts exist?

Within the context of evolutionary theory, the arts have no obvious natural selection value...

...but a predisposition to model and evaluate possibilities before they occur, and to investigate preferred possibilities, would have such value...

A puzzle for evolution: Why do the arts exist?

Within the context of evolutionary theory, the arts have no obvious natural selection value...

...but a predisposition to model and evaluate possibilities before they occur, and to investigate preferred possibilities, would have such value...

...especially if the possibilities included not only novel actions, but also novel artefacts...

A puzzle for evolution: Why do the arts exist?

Within the context of evolutionary theory, the arts have no obvious natural selection value...

...but a predisposition to model and evaluate possibilities before they occur, and to investigate preferred possibilities, would have such value...

...especially if the possibilities included not only novel actions, but also novel artefacts...

...and if the investigation of the preferred possibilities included the construction of the imagined artefacts

A puzzle for evolution: Why do the arts exist?

Will our robot spontaneously perform some non-mission related action or modify its environment to make it 'feel good'?

A puzzle for evolution: Why do the arts exist?

Will our robot spontaneously perform some non-mission related action or modify its environment to make it 'feel good'?

Will it do so after simulating and evaluating that action or modification?

A puzzle for evolution: Why do the arts exist?

Will our robot spontaneously perform some non-mission related action or modify its environment to make it 'feel good'?

Will it do so after simulating and evaluating that action or modification?

And finally, If it does, will this show that an architecture for enabling intelligence necessarily produces aesthetically-driven creativity?

For more information see

www.machineconsciousness.org

And please feel free to email me

owen@essex.ac.uk