The Construction of Light

Ron Chrisley

Sackler Centre for Consciousness Science
Centre for Research in Cognitive Science
PAICS Lab
School of Engineering and Informatics
University of Sussex

Beyond AI 2013
November 12-14, Plzen, Czech Republic

PAICS

Sackler Centre for
Consciousness Science

US University
of Sussex
Informatics

COGS
Centre for Research in Cognitive Science

0. Talk Talk

Business as Usual?

- (My) talks are usually *narrowcast*:
  - One topic
  - In-depth
  - Terms carefully defined
  - Arguments
- Thus, often of interest only to a few people in the audience
- But then I saw this:

BEYOND AI
ARTIFICIAL GOLEM INTELLIGENCE 2013
november 12-14, 2013
Pilsen, University of West Bohemia

1. ***History of AGI***
   - AGI as the original goal of AI research
   - pursuit of cybernetic brain
   - AI and cybernetics
   - dissolution of AI into engineering
   - AGI rise & fall & rise
2. ***Nature of Intelligence***
   - alternatives to AGI
   - natural and artificial intelligence
   - humans as AGI
   - externalism and extended mind
   - whole-brain emulation
3. ***Risks and Ethical Challenges***
   - cognitive biases
   - switching off AGI
   - AGIs as slaves
   - responsibility of AGI's agency
4. ***Faith in AGI***
   - AGI and Genesis 1:26
   - Rabbi Loew & AGI & Golem
   - Shemhamphorasch *in silico*
   - Kabbalah in AI labs
5. ***Social and Cultural Discourse***
   - us & the others
   - gender in humans and AGIs
   - social and narrative construction of AGI
   - science fiction as social reality
   - cross-cultural perception
6. ***Invoking Emergence***
   - multiagent systems
   - neural networks as the second best way to implement a solution
   - is GOFAI dead?
7. ***AI and Art***
   - cyborg and robot design
   - uncanny valley
   - cyberpunk art
   - machines as artists

"I suggest a new strategy, R2…"

- The fascinating, wide-ranging remit of this conference prompted a change of strategy
- This talk will be *broadcast*:
  - Multiple topics
  - Briefly covered
  - Terms/references often not explained
  - Hand waving

Broadcasting

- More likely to generate interest with more of you
- Hopefully prompting detailed discussion “off-line”: coffee breaks, lunch, dinner, twitter, email…
- So don’t tune out: even if one slide bores/baffles you, the next one might be of interest
- For the first few sections, I highlight in green the connections to the remit of the conference

BEYOND AI
ARTIFICIAL GOLEM INTELLIGENCE 2013

november 12-14, 2013
Pilsen, University of West Bohemia

1. ***History of AGI***
   - AGI as the original goal of AI research
   - pursuit of cybernetic brain
   - AI and cybernetics
   - dissolution of AI into engineering
   - AGI rise & fall & rise
2. ***Nature of Intelligence***
   - alternatives to AGI
   - natural and artificial intelligence
   - humans as AGI
   - externalism and extended mind
   - whole-brain emulation
3. ***Risks and Ethical Challenges***
   - cognitive biases
   - switching off AGI
   - AGIs as slaves
   - responsibility of AGI's agency
4. ***Faith in AGI***
   - AGI and Genesis 1:26
   - Rabbi Loew & AGI & Golem
   - Shemhamphorasch *in silico*
   - Kabbalah in AI labs
5. ***Social and Cultural Discourse***
   - us & the others
   - gender in humans and AGIs
   - social and narrative construction of AGI
   - science fiction as social reality
   - cross-cultural perception
6. ***Invoking Emergence***
   - multiagent systems
   - neural networks as the second best way to implement a solution
   - is GOFAI dead?
7. ***AI and Art***
   - cyborg and robot design
   - uncanny valley
   - cyberpunk art
   - machines as artists

BEYOND AI
ARTIFICIAL GOLEM INTELLIGENCE 2013
november 12-14, 2013
Pilsen, University of West Bohemia

1. *History of AGI*

- AGI as the original goal of AI research
- pursuit of cybernetic brain
- AI and cybernetics
- dissolution of AI into engineering
- AGI rise & fall & rise

2. ***Nature of Intelligence***

- alternatives to AGI
- natural and artificial intelligence
- humans as AGI
- externalism and extended mind
- whole-brain emulation

3. ***Risks and Ethical Challenges***

- cognitive biases
- switching off AGI
- AGIs as slaves
- responsibility of AGI's agency

4. ***Faith in AGI***

- AGI and Genesis 1:26
- Rabbi Loew & AGI & Golem
- Shemhamphorasch *in silico*
- Kabbalah in AI labs

5. ***Social and Cultural Discourse***

- us & the others
- gender in humans and AGIs
- social and narrative construction of AGI
- science fiction as social reality
- cross-cultural perception

6. ***Invoking Emergence***

- multiagent systems
- neural networks as the second best way to implement a solution
- is GOFAI dead?

7. ***AI and Art***

- cyborg and robot design
- uncanny valley
- cyberpunk art
- machines as artists

1: The Nature of the Artificial

What is AGI, anyway?

- Or rather, which notions of AGI are interesting, problematic, potentially desirable?
- Trivialization threat:
  - AGI as "machines that think" implies (even to Searle) that we are AGI, that making babies is doing AGI, etc. (humans as AGI)
- Two approaches:
  - Emphasis on substrate: mind realised in different material?
    - Perhaps, but would we really think AGI to be achieved if we discovered that some of us have different insides?
    - Actually we do! How different is different enough?
  - Robust notion of the artificial
    - Not just caused, but *designed* by us (natural and artificial AGI)

Creation and AGI

- On Judeo-Christian views, would this mean Adam was the first AGI? (AGI and Genesis 1:26)
  - Unlike Christ, we are told, we were made by, not begotten of, God: precisely the causing vs designing distinction
- But the proper dichotomy is not natural vs. *artificial*; it is natural vs. *super-natural*
- This allows the artificial to be a subclass of the natural
- Incidentally: AGI is sometimes thought to be blasphemous (hubris far beyond the Tower of Babel)
  - But not so if we were made *imago dei*: If God both begets and makes minds, then we ought to as well?

Theodicy and AGI

- Actually, God has a famous problem: How can He be omnibenevolent, omniscient, and omnipotent, given that there is evil in the world?
  - I.e., Theodicy: Isn't God responsible for (our) evil?
  - How can God make (not beget) us, and yet not be the author of our actions?
  - AGI researchers have the same problem: the paradox of AGI:
    - Inasmuch as a system is designed by us, we, not it, are responsible for its actions; thus it is not a true mind (AGIs as slaves)
    - Inasmuch as a system is not designed by us, it is not artificial
  - Not just a theoretical problem: Many of GOFAI's difficulties stemmed from a too-close connection between the designer's and designed's registrations of the world (is GOFAI dead?)

AGI and Theodicy

- AI developed strategies for dealing with this: adaptivity
  - Learning, but even more: artificial evolution, genetic algorithms
- Can these solutions be transferred to theology?
- High irony: Far from being a threat to God's role as creator, evolution becomes the only way to solve problems of Theodicy (Faith in AGI)

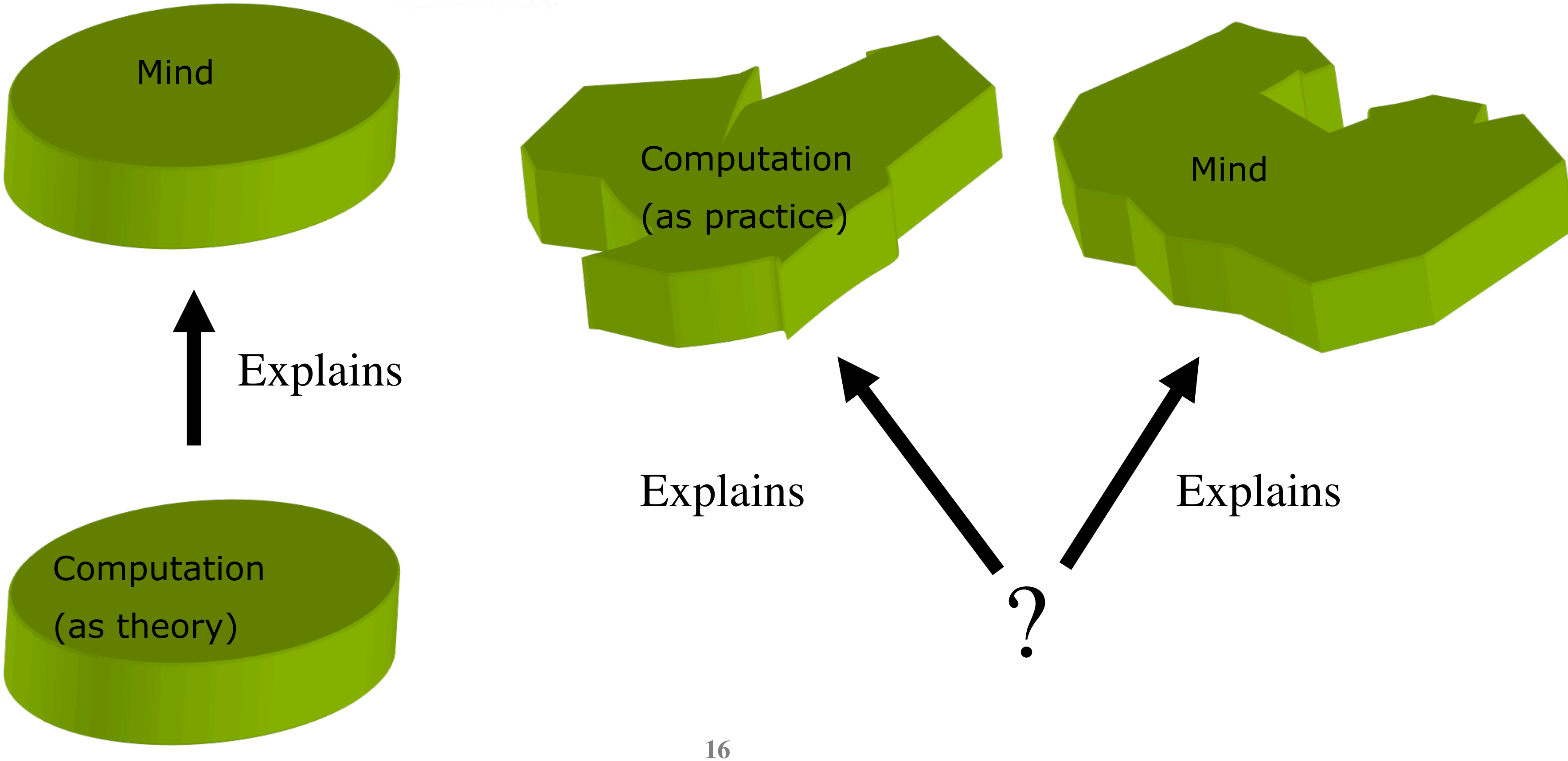
- (A further consideration: can the philosophical tools developed by cognitive science for naturalizing the mind also make clear the conditions for naturalizing God and the spiritual?)

2: A Role for Computation

Personal backdrop

- My interest is not just A(G)I, but Cognitive Science: computation and mind
- Aware of the criticisms, but a reformer, not a revolutionary
- One source of trouble is not Genesis 1:26, but John 1:1 – "In the beginning was the word" (Faith in AGI)
- That is, an overly linguo-centric approach has limited and misled AI and Cognitive Science (and their critics) on many occasions, e.g.:
    - The "Turing test"
    - Diagonal/Gödelian critiques of AGI
    - Jackson's Mary/Knowledge Argument against physicalism
    - Symbolic A(G)I/The Language of Thought (is GOFAI dead?)
    - Over-emphasis on conceptual content

Misconstruing Computational Explanation 1



Mind

Explains

Computation (as theory)

Computation (as practice)

Mind

Explains

Explains

?

Misconstruing Computational Explanation 1

- In particular, many of the criticisms of a computational explanation of mind have the form: Mind is X but computation is not
  - Where X = extended, embodied, semantic, dynamic, socially mediated, etc.
- Usually that is based on a narrow, theory-driven view of what computation is
- If you look at computation as it is practiced “in the wild” (Cantwell Smith), you’ll see that it is all those things too (Agre)
- In short, computation doesn’t explain, e.g., intentionality
- But perhaps (part of) a good way to develop the theories, concepts and tools required to explain human intentionality is to try to first explain the intentionality of computers

The Extended Mind?

Yes, externalism is interesting and true, but:

- One must be clear about the proper notion of external (hint: it's a phenomenal, not a spatial notion)
- The Parity Principle is not helpful
- The examples Clark and Chalmers give are not cases of extended minds:
  - Otto's beliefs are not in his notebook
  - Otto's case is not analogous to Inga's: Otto's notebook is only involved because it is first an object of Otto's experience; not so for Inga's neurons.

Transparency and Prosthesis

- True mind extension has to involve true transparency: implants/prostheses (alternatives to AGI)
- This has implications for the proper treatment of sensory augmentation devices (e.g., which sensory modality do they provide?)
- Prosthetic approaches are under-appreciated in (the philosophy of) A(G)I
- Will say more in section on machine consciousness

A Mereological Constraint

- Back to the difference between Otto and Inga
- Proposal – The Mereological Constraint:
  - Strong version: A system X+Y can't count as a mind if X already counts as a mind
  - Weak version: A system X+Y can't count as a mind if X already counts as a mind, and if X's being a mind explains why X and Y form a system in the first place
- Even the weak version would be enough to exclude Otto as a case of extended mind
- But would also deflect the Chinese room (Searle) and Chinese nation (Block) arguments against AGI (is GOFAI dead?)

Misconstruing Computational Explanation 2

- But there is another way to defeat the Chinese room argument, which clarifies the enterprise of AGI and MC
- Searle assumes that if computation isn't Strong AI = *sufficient* for mind (understanding), then it is Weak AI = merely a tool, as it is in meteorology
- But there is a middle possibility: the computational properties investigated by A(G)I are *necessary* for mind
- This makes computation explanatorily essential to understanding mind
- It also enables a form of the robot reply: yes, you need more than programsfor AGI, but computers and robots comprise more than programs (cf Misconstruing Computational Explanation 1)

The Middle Way

Put another way:

- Yes, programs aren’t in themselves enough to explain minds…
  - …but neither are they alone enough to explain computers!
- Further, that still leaves the possibility that, as with computers, they are often an essential (even dominant) part of an explanation

Is Computation Real?

- As an aside, the problem with objections (originally from Putnam and Searle, but now also from, e.g., Bishop and Gamez) that computation is observer-relative and so cannot explain minds…
- …is that they also imply that computation cannot explain computers
- They also fail to grasp that in essence, computational vehicles are *counterfactual*
- Occurrent computational states don't just involve occurrent physical states, but also the physical states the system would go into were it to receive this or that input
- Thus, well-suited in fact to be vehicles for expectational or predictive coding accounts of experience

3. Machine Consciousness

Artificial Qualia

- This middle way (computation as necessary, if not sufficient, for mind) is especially helpful in making room for MC
- But one stumbling block for many is qualia (the "what it is like"-ness of experience):
  - Ineffable
  - Intrinsic
  - Immediate
  - Private
- How could a computational (or indeed any physicalist) approach account for them?
- Dennett: It can't, so eliminate them

Ontologically Conservative Heterophenomenology

- But perhaps qualia are like gold
- Ancients/mediaevals had a false theory of gold (gold is not a compound, let alone one with phlogiston as a constituent)
- We now know that nothing meets their definition…
- …but we don't say that we have shown that gold does not exist!
- Rather, we have given a better account of the phenomenon that prompted the ancient, inaccurate account

Ontologically Conservative Heterophenomenology

- Perhaps the same can happen for qualia: what are the systemic properties that might prompt a robot to take itself to have intrinsic, immediate, ineffable, private states?
- There may or may not be a unified account that underwrites ontological conservation: empirical investigation required
- So Dennett should be more thorough-going in his empiricism: It is as wrong to *a priori* eliminate qualia as it is to *a priori* assert them

What About the Hard Problem?

- What happened with gold between ancient and modern times was: conceptual change, referential stasis
- Similarly, suppose it is true that on our current conception of consciousness, there is a hard problem: zombies are possible
- Might there be a concept of the same phenomenon that does not carry with it the possibility of zombies?

Interactive Empiricism

- Plausibly, we cannot get to such a concept through conceptual reasoning alone
  - Wittgenstein: “seeing as” is mastery of a technique
  - Mastery of a technique requires practice: experiential activity
  - Cf Held & Hein’s cats

Interactive Empiricism

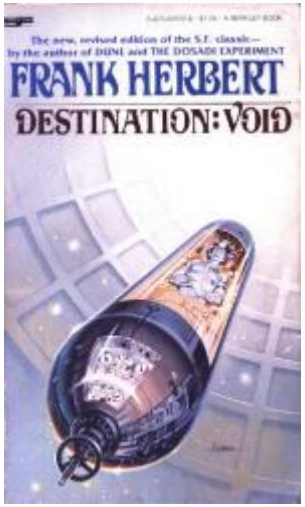
- Suggests a revision to our notion of philosophy: Sometimes conceptual problems are solved, not by working with the conceptual primitives we have, but by changing/adding to (and subtracting from?) the stock of primitives itself.
  - Philosophy is sometimes experiential activity, not just armchair analysis
  - And so is science (which explains why Jackson's Mary doesn't have all the physical information if she hasn't seen red)

Interactive Empiricism: We Are Part of the MC System

- Engaging in machine consciousness research (and interacting with its products) might provide the kind of experiential activity to prompt conceptual change
- Example: The enactive torch
- Our design should reflect this, in two ways
  - Acknowledging our interactive role in machine development (Kismet's eyebrows)
  - Engineering for our conceptual change

Frank Herbert's Prescience

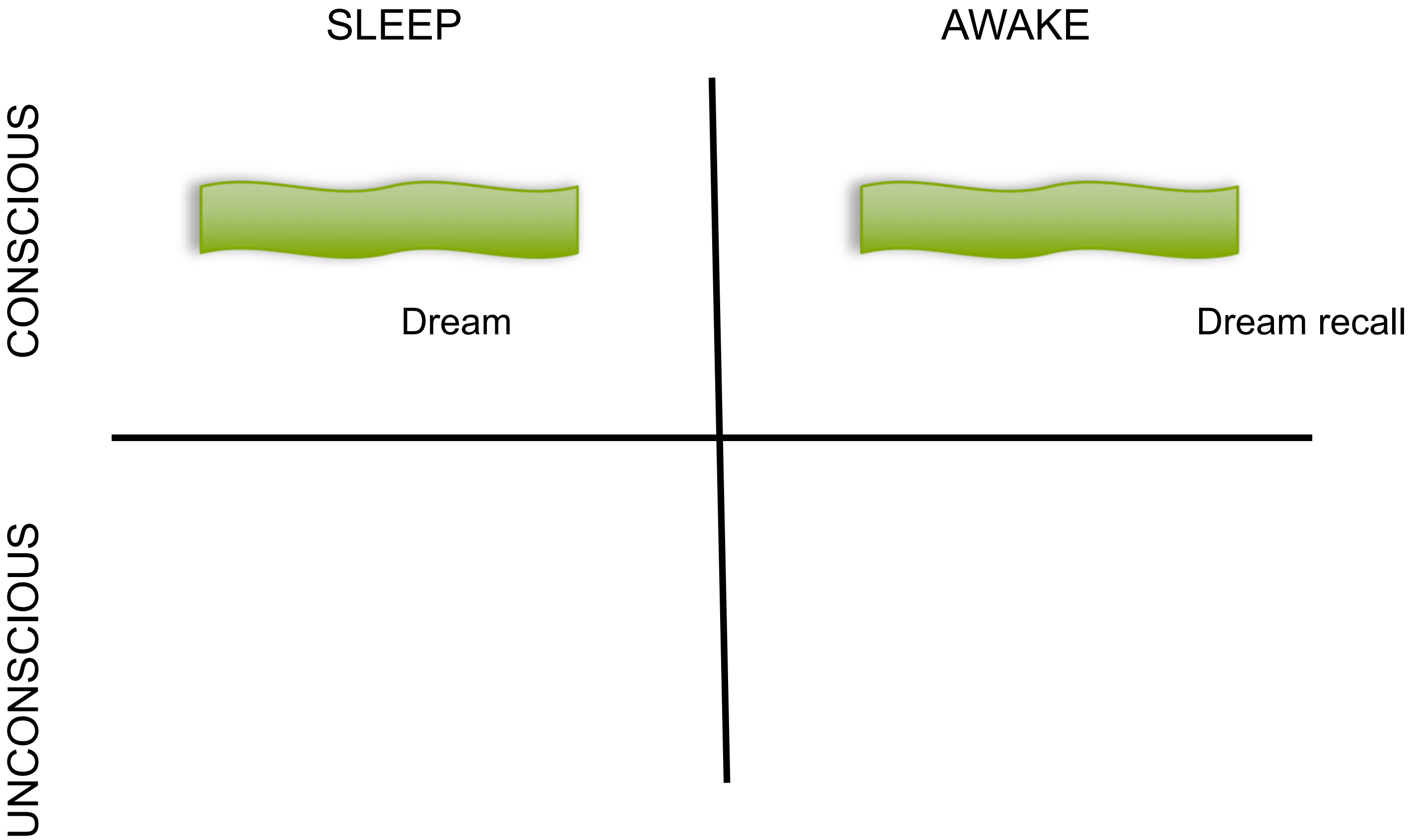
- In the science fiction novel *Destination: Void*, scientists who want to create machine consciousness decide that the best way to do it is not via direct engineering, but to design a situation in which:
  - Carefully engineered people (clones)
  - Are sent into deep space in a carefully engineered technological environment (computers, neural wetware)
  - Are manipulated and motivated to find a way to create machine consciousness (e.g., they will die if they don't!)
  - A crucial part of the project is for the challenges they face and the technology they build to play a role in them figuring out what consciousness *is* (conceptual change!)


The new, revised edition of the S.F. classic—
by the author of DUNE and THE DOSADI EXPERIMENT
FRANK HERBERT
DESTINATION: VOID

MC and Dreams

- What can dreams tell us about consciousness?
- Might a better understanding of, e.g., the role of narrative construction and sense-making in dreams give us suggestions for how to achieve MC?

(Naive) Realism



SLEEP

AWAKE

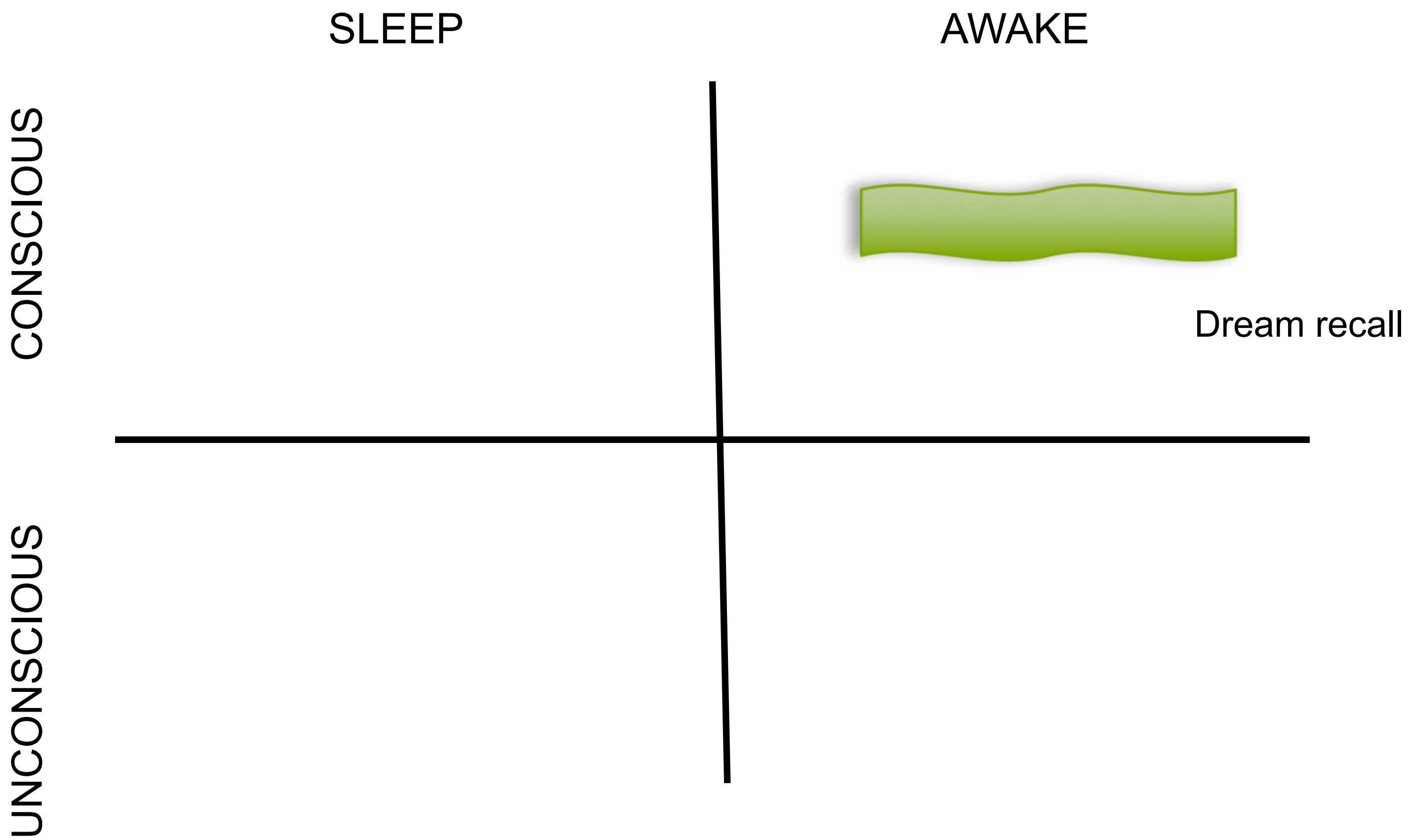
CONSCIOUS

UNCONSCIOUS

Dream

Dream recall

(Extreme) Anti-Realism



SLEEP

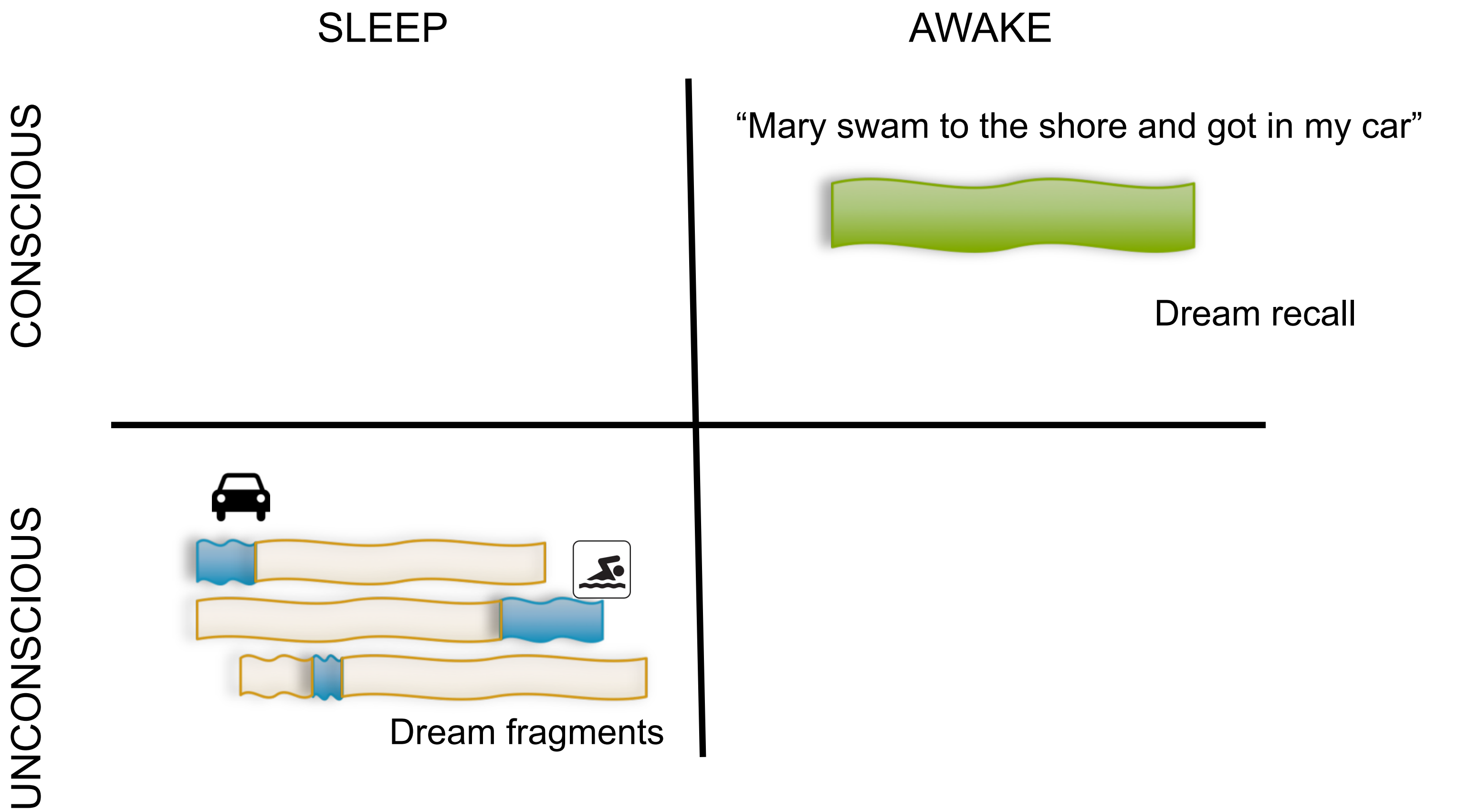
AWAKE

CONSCIOUS

UNCONSCIOUS

Dream recall

Moderate Anti-Realism



SLEEP

AWAKE

CONSCIOUS

UNCONSCIOUS

“Mary swam to the shore and got in my car”

Dream recall

Dream fragments

So Also For Conscious Experiences?

- The moderate anti-realist view could explain paradoxical temporal experiential phenomena such as:
  - Libet's subjective referral in time
  - Colour-phi
  - The cutaneous rabbit
  - Flash-lag

The Prosthetic Alternative Revisited

- Prosthetic machine consciousness (MC) is in a good position to side-step some standard objections, such as:
  - How could you know if your MC is actually conscious?
  - How could you know which conscious states it is in?
  - Didn't Searle show that only biological systems have the causal powers necessary for consciousness?
  - Etc.
- Meanwhile, systematic exploration of the interdependence between the phenomenal and the physical could occur
- Like interventionist neuroscience (e.g., TMS) + auto-cerebroscopes, but substrate-neutral, prompting a more general account

Content Specification

- Concepts are components of representational content that are:
  - Arbitrarily recombinable
  - Under endogenous control
  - Such that the question of justification arises
  - Articulable
- Conceptual contents can thus be precisely specified by expressing them in language
- Part of overcoming the linguo-centrism of GOFAI is the acknowledgement of non-conceptual experiential states
- Unlike conceptual contents, not, in general, expressible in language
- So how are we to specify, e.g., the non-conceptual contents of visual experience?

Virtual Reality Specifications

- A picture is worth a thousand words; better: video
- Even better: virtual reality
  - Specify experiences to each other by giving parameters for reference VR systems: "The experience you have when you use the VR system loaded with these parameters"
  - Specifies by exploiting the experiential capacities of the user of the specification
  - Shows again how Jackson's thought experiment assumes a contradiction – Mary has all physical info but has never seen red?

Synthetic Phenomenology

- Another approach: use working AGI/MC robotic models of, e.g., visual experience, as reference systems
- Specify experiences by referring to the states/properties (both internal and external/relational) of the robot when it is modeling a particular experience
- A structureless listing of these properties would be unhelpful
- Rather, these states can de depicted in a way that (again) exploits the experiential capacities of the person using the specification ("enactive depictions")
- E.g. SEER-3 robot and expectation-based qualitative theory of experience

4. Ethics and AGI

Ethical Zombies vs. Ethical Zombies

- Work in the growing field of machine ethics looks set to make some of the same mistakes as (the philosophy of) AI and MC
- Specifically: taking seriously the possibility of physical/functional/behavioural duplicates that nonetheless differ in their ethical status (agents or patients)
- Can tolerate behavioural duplicates with different cognitive or phenomenal properties
  - The Turing test not withstanding, AI/CogSci is about “internal” states/processes, not behaviour alone
- But should not view behavioural duplicates as possibly ethically different

Ethical Zombies vs. Ethical Zombies

- If we allow there to be any criterion for ethical status beyond the behavioural, we open the door to atrocity
  - For all I know, I do not meet that criterion!
- Should not set the bar for artifacts counting as ethical patients so high that humans could not pass it
- But this behaviourism about ethical status does not sit well with "internalism" about consciousness (cf my Mereological Constraint)

Ethical Zombies vs. Ethical Zombies

- So much the worse for consciousness being the foundation of ethics!
- Already have good reasons for that, anyway:
  - Corporations
  - Infants/foetuses
  - Unconscious subjects
  - Some animals

5. AI and Art

(in one slide)

Nine Principles for Artificial Artistic Creativity

1. If you make your robot pleasure-seeking, and make creativity pleasurable, you'll make your robot creative
2. To be a good creator, it helps to be an appreciator
3. Let the robot experience output in the real world, as we do
4. We won’t like what it likes unless it likes what we like
5. An important motivator is the approval or attention of others
6. Sometimes it is better not to try pursue novelty directly, but something that is correlated with it: the subjective edge of chaos
7. Let dynamics play a role in appreciation
8. Patterns in one's own states can be the objects of appreciation
9. The best way to make outputs in the real world is to be embodied in the real world

And if a bird can speak, who once was a dinosaur
And a dog can dream; should it be implausible
That a man might supervise
The construction of light?

– Adrian Belew

Thank you.

Comments welcome:

ronc@sussex.ac.uk

Twitter: @ronchrisley