Artificial Intelligence and Consciousness.

Yorick Wilks
Florida Institute of Human and Machine Cognition
and
Internet Institute, University of Oxford
www.dcs.shef.ac.uk/~yorick

Gresham College, London, January 2020

Information Society
Technologies

What the talk is about

- What is consciousness and why is it a philosophical problem, as it is now seen to be?
- What has consciousness do with AI?
- Could AI entities be conscious?
- How would we go about making them so?
- Would it matter if they were or were not?
- How would we *know* if they were?

Starting point: Radical differences about what consciousness *is* in the early C20.

- William James : "'consciousness' is the name of a non-entity, and has no right to a place among first principles"
- Sigmund Freud: "What is meant by 'consciousness' we need not discuss. It is beyond all doubt"

Consciousness is not a *traditional* philosophical problem

- It does not feature in traditional or classical philosophy at all.
- Until recently it was not possible to discuss it within the Anglophone analytic tradition.
- Yet now David Chalmers says “it is *the* hard problem”. Why?
- Chalmer’s position: that even a complete specification of a creature in physical terms leaves unanswered the question of whether or not the creature is conscious.

Is consciousness a historical phenomenon?

- Were Shakespeare and Plato as conscious as us?—they didn’t mention it.
- German 19C and the notion of *Bewusstsein*
- Jaynes theory of the historical origin of consciousness and the Old Testament prophets---its relation to language?

Chalmers “hard problem”

- It is not (he says) an easy problem like perception or planning—which can be explained by physiology and causes in the brain
- But the *“state of being like this”*
- Some have called it “what its like to be us” (as opposed to “being a bat”) and used the term “qualia” for it.

Chalmers thinks consciousness not explicable fully by brain causes

- He doesn’t deny that consciouness does rest on brain machinery only that that isnt sufficient to explain its content and nature.

Trasitional Philosophical starting point

- DUALISM: DesCartes C17 doctrine that mind and matter are different substances that interact somehow.

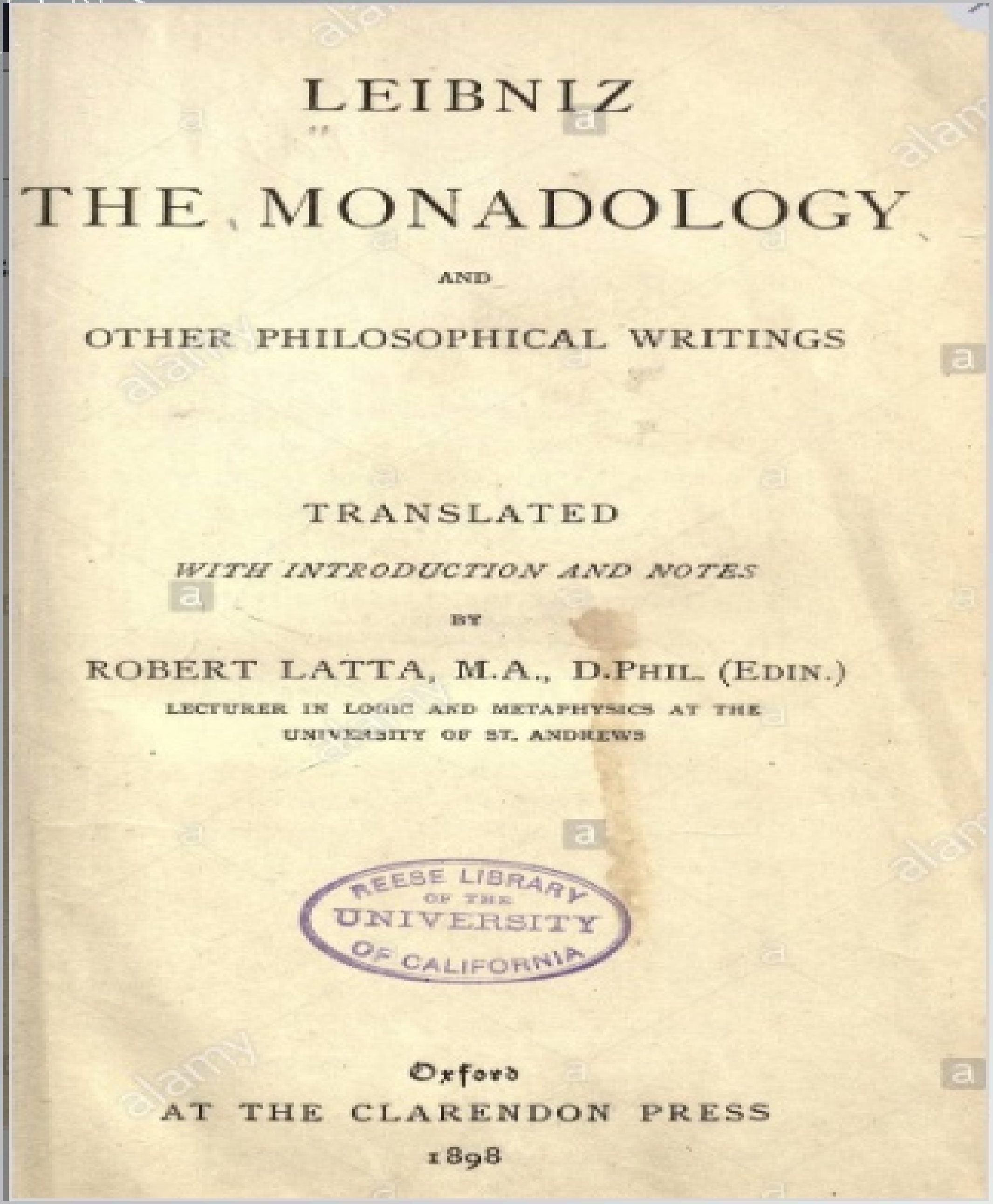
Not clear that consciousness is different from a perception, at least when Leibniz used the term

- Leibniz (17C) talked of entering a mill and seeing its works
- He also supposed we could enter a machine that perceived and reasoned
- But we would not then see its *perceptions* as we examined it.
- Is that the same question as “what is consciousness”?

Leibniz’s words (translated)

- “It must be confessed, moreover, that perception, and that which depends on it, are inexplicable by mechanical causes, that is, by figures and motions, And, supposing that there were a mechanism so constructed as to think, feel and have perception, we might enter it as into a mill. And this granted, we should only find on visiting it, pieces which push one against another, but never anything by which to explain a perception. This must be sought, therefore, ……… not in the composite or in the machine.”

- If that means “how we see” then we do now have an explanation, but not if it means consciousness.



LEIBNIZ

THE MONADOLOGY

AND

OTHER PHILOSOPHICAL WRITINGS

TRANSLATED

WITH INTRODUCTION AND NOTES

BY

ROBERT LATTA, M.A., D.PHIL. (EDIN.)

LECTURER IN LOGIC AND METAPHYSICS AT THE
UNIVERSITY OF ST. ANDREWS

Oxford
AT THE CLARENDON PRESS
1898

"Perhaps the cleverest man who has ever lived…." Bertrand Russell

Gottfried Wilhelm Leibniz.

Leibniz is a bridge to AI

- He wanted a *reasoning machine* and
- An artificial language for reasoning in.
- The *Monadology*: all things are conscious in their own way, some dimmer than others (and only God is fully conscious of everything).
- The idea that *all things are conscious*—we shall return to that later.

A huge change in Anglosphere philosophy

- Francis Crick, discoverer of DNA, recommended never to mention the term "consciousness" in a grant application, or it would be refused.
- Until 1980s the necessarily private was undiscussable in Anglosphere philosophy
- Behaviourism, Ryle, Malcolm on dreams, Wittgenstein on pains and a private language
- Only Continentals talked about experience and consciousness, and consciousness being *attention to something*.

Meanwhile, in the EU…

- Panpsychism: Aristotle and Leibniz: Everything is conscious its degree
- Hegel’s C19 *Bewusstsein* and the whole world becoming conscious with the “human layer”
- This is a non-individual consciousness, no privacy, knowing ourselves through others
- Or, The whole world is a single conscious thing (Spinoza C17)
- Teilhard de Chardin and Jung’s “collective unconscious”.

Panpsychism revived

- A new form of idealism: where mental phenomena are the real ones,
- This is not DesCartes dualism (mind AND matter) or individual idealism (there’s just me and *my* mind said Berkeley)
- Consciousness IS the stuff of reality, right down to electrons
- It is also form of physicalism—there is only matter BUT it is conscious
- physical reality cannot be strictly separated from the mind (cf. quanta)

The Anglo/Brexit anti-consciousness opposition: Daniel Dennett

- Dennett thinks consciousness is basically empty
- There is nothing there really (Compare Hume C18 on the non-existent self).
- Consciouness (he says) is like a Public Relations official who is handed a sheet with what to say on it, but has no part in its creation.

What is real and central

- Dennett denies the reality of experience, Galen Strawson makes it central and the most real thing.
- Strawson: consciousness is “the only thing in the universe whose ultimate intrinsic nature we can claim to know.”

Early ideas that our conscious minds don’t really know whats going on “underneath”;

- George Eliot’s Adam Bede (1859):
- "Our mental business is carried on in much the same way as the business of the state: a great deal of hard work is done by agents who are not acknowledged".
- And, of course, Freud’s Unconscious.
- Dennett, and much AI (as we shall see).

An interesting question from Chalmers: there could be Zombies?

- Just like us but *not conscious at all*
- How could we tell—what could we ask them ?
- What difference would it make to them?
- And I don't know about *your* consciousness do I?
- You cant tell just by looking and listening (dogs? Just as alert as some folk?)

Would there be moral implications for conscious machines?

- Suppose machines were *not* zombies.
- Many feel a link between our treatment of animals and the fear they might be conscious.
- If industrial robots were conscious would this revive a Marxism with modern automation ----for a new exploited proletariat?

Could there be different types of consciousness in people; though it’s hard for any one of us to know that?

- People see colour differently (colour blindness)
- Some people cannot see optical illusions
- Split brain patients are “aware” of things but do not know they are.

Where does AI come in?

- Classic questions: could a machine be conscious and how would we know it it was?
- What kind of progamming/theory would we perform or create to make it so (Answer: we don’t know!!).
- Would AI entities be better if they were conscious, or should they stay zombies?

The distinctive AI idea about consciousness is from Minsky

- An attention mechanism---so we don’t need to know what lower levels of our body and mind machines are doing (breathing, digesting…how could we know all those things?)
- But maybe Gurus can do that?
- attention=consciousness is by definition restrictive and partial

In AI this idea has been associated with "levels" of software and programming languages and closed modules

- Levels of programming languages: at the top is English: if you tell a driverless car "Take me to Wigan".
- Also modules interacting (Hewitt): all are "black boxes" with no central control?
- Both can be "Dennettish" in people—depending on where the *real* decisions are made.

Two ways of looking at the "top level" of control in humans and machines

- Which we can identify with consciousness
- We don't know the "lower levels" of how things work and that's essential to real control, meaning, intention (eg how my arm works)…..(=> Minsky)
- We don't know the "lower levels" of how things work *and where decisions are made* so the top conscious level is vacuous (=>Dennett)

The opaqueness of programming language levels works in both directions

- *Downwards:* the top level language code does not have access to how its commands are carried out at lower levels.
- *Upwards"* the top level instruction—what it was "really doing"-- cannot be decoded from the lower level code, so a brain's conscious content could not be just decoded from its neurons.

The neural network AI paradigm

- Huge networks that learn and whose function is not wholly understood.
- This has given rise to an “Emergence” theory of consciousness
- Where *any* sufficiently complex network will become conscious.
- Neurophysiologist Graziano: a woman who has lost an arm thinks it’s still there, like a phantom limb: “One is the ghost in the body and the other (consciousness) is the ghost in the head.”

Could the WWW be Hegel's world-wide spirit of humanity—the conscious layer of the universe?

- William Gibson in NEUROMANCER 1981 invented the term "cyberspace":
- Cyberspace. A consensual hallucination experienced daily by billions of legitimate operators, in every nation. . . . A graphic representation of data abstracted from the banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding.

Consequences for AI of any form of panpsychism

- If “everything is conscious”, then machines will be conscious too, as part of everything
- Japanese attitudes to robots and “world spirits”—different from the Western attitude
- But, that isn’t what is usually meant by “machine consciousness”
- Not like all objects *but like us!*

AI-consciousness links following Minsky's line on top-level control

- Neil Lawrence's theory of information transfer rates in humans (language) and machines (data); requires knowledge structures and models of others.
- Link back to phenomenology/Husserl consciousness was always OF something

Links to evolution of language and machines that talk to themselves

- Jaynes on prophets, language and self-consciousness.
- Lawrence’s information transfer distinction leads to complex AI models and Minsky-like control of alternatives.
- Self-discussion of plans as central to consciousness.
- Cleermans on machines *learning to be conscious* via constant redescription of their own activity.

Consciousness and intentional action

- Bello has argued that intentional action requires consciousness in the sense of “paying attention” to a purpose
- Story of killing the uncle with a car while not paying attention, even though intending to kill him later.
- This argues against Zombies with intelligence but no consciousness

Could there be experiments to detect/determine if an AI entity is conscious?

- Could a machine have the privacy/authority we have over its own function?
- Other speculations on complex mental manipulations like imagining being “out of the body” and having consciousness.
- A Templeton project to determine, via brain electrical activity, whether *Global Workspace Theory* or *Integrated Information Theory* is true of humans.
- First is front-brain control, latter is back brain.

Yampolsky: a (semi-Turing) test for AI consciousness

- You present an agent with a novel optical illusion and a set of choices.
- The agent answers in a such a way as to suggest it “gets”/sees the illusion (like the duck AND the rabbit—one the team understands but they *must* be novel)
- Same for a machine.
- Inference to “X is having the experience I am” –though without me having the central “inner” experience, making it non private.

Takeaway thoughts

- AI may not need consciousness, but may get it if we could work out how we knew we had succeeded in creating it.
- Its existence in machines is almost certainly tied to “self-conversation”
- But there would then be new moral and social problems.