The Philosophy and Ethics of Artificial Intelligence

(Artificial) Consciousness

Overview of Course

- Introduction
- Intelligence
- Consciousness
- Reasoning and Communication
- Ethics and Morality
- Algorithms
- For good or bad ? AI in Medicine, AI in War
- Superintelligence and the Value Loading Problem
- AI and Human Society
- Review/Revision

(Artificial) Consciousness

Overview of Today's Lecture

- What is Consciousness ?
- Problems of Consciousness
- Theories of Consciousness
- Could an AI be conscious ?
- Implications

Sources and Reading

- Life 3.0. Max Tegmark
- Stanford Encyclopedia of Philosophy (https://plato.stanford.edu/index.html)
- Hedda Hassel Mørch asks: what is IIT all about?
- and others ….

Consciousness

Consciousness is something with which we're all intimately familiar. It's the thing that goes away every night in deep sleep, and comes back when we wake up every morning, or whenever we start dreaming. It encompasses all our subjective feelings and experiences, ranging from the simple redness of red…**It's the one thing that is directly and immediately known to us**, and it mediates our knowledge of the external world.

Giulio Tononi (neuroscientist)

And yet it is one of the most profound and seemingly insoluble mysteries of the universe

Revisiting the Chinese Room

The Chinese Room Continued

- Recall assumption in Chinese room argument

  intelligence ➔ understanding

- But as we've already discussed, is "understanding"

  in the sense that there is an explicit ***conscious*** *awareness/apprehension* of the relationship between the Chinese symbols and the meaning of these symbols

  necessary for intelligence ?

"Competence without Comprehension" – Daniel Dennet

- After all, if words can be used in all the ways that communicate the right kind of information (algorithms) for achieving complex goals, then why do we need this extra internal apprehension of what they mean ?

- Consider that in modern NLP computers can respond to questions, and answer them (IBM Watson and more recently GPT-3), based on learnt contexts (sentences) in which words appear – the meaning of words are the totality of contexts in which they are used

  "a word is characterized by the company it keeps"
  "the meaning of a word is all the ways the word can be used"

Arguably then, isn't the Chinese room really only an argument against the possibility that machines can have conscious experiences – what *it feels like to* understand something ?

Understanding is as Understanding does

- To conclude: Chinese room is an argument against AGI if one accepts that **a)** understanding is necessary for intelligence and **b)** that understanding must include some inner awareness or feeling that accompanies understanding. But

- one can deny a) (as mentioned in the last lecture) and/or
- one can deny b) (as illustrated by modern techniques in NLP)

- After all, the only way to recognise that some entity (machine or human) understands something is by observing that the entity can do everything that would be implied by such an understanding

- There may not be anything going on in the "head" of a machine that is the **feeling** (the conscious attendant) of understanding, but the machine could be seen to take advantage of what it is to understand something in all the ways that one might imagine what it would mean to take advantage of what it is to understand something

The Chinese Room Continued

- Indeed Searle restated the conclusion of his argument 30 years later in 2010:

"I demonstrated years ago with the so-called Chinese Room Argument that the implementation of the computer program is not by itself sufficient for ***consciousness or intentionality*** "

Intentionality = "aboutness"

An implication of this revised interpretation:

A computer may behave (by virtue of the replies it gives) ***as if*** there were ***real understanding*** (i.e., that it *consciously* understood what was being asked and what it was saying), but in fact there is no "real" understanding

But again, I ask you, if a machine can do everything/behave as if there were real understanding, why not say that the machine has "real" understanding

The Chinese Room Continued

Searle is arguing against ***functionalism*** :

the view that states/ processes count as being of a given mental or conscious type because of the functional role they play

(if it looks like, walks and quacks like a duck, it is a duck !)

(if it behaves as if it has real understanding, then it has real understanding)

Functionalism as applied to conscious experiences in general, and not just the conscious feelings that accompany understanding:

If a system S1 produces the same input/output pairs as a system S2 that we know to be conscious, then SI is conscious (and whether S2 is made from the same stuff as S1 doesn't matter)

The Chinese Room Continued

Searle's main target is a form of functionalism known as the Computational Theory of Mind that treats minds as *information processing systems*.

According to functionalism an artificially intelligent machine is conscious if it behaved (functioned) ***as if*** it were conscious (it it walks, talks and looks like a duck, it is a duck)

What is Consciousness ?

Consciousness = subjective experience

(Life 3.0. Max Tegmark)

Something is conscious if there is something that it is like to be that something, i.e., there is some subjective way the world seems or appears from the creature's mental/experiential point of view

"What Is It Like to Be a Bat?". The Philosophical Review. 83 (4): 435–45. Thomas Nagel (1974).

- Consciousness is the most important and arguably mysterious feature of the universe.

  Without consciousness there can be no happiness, goodness, beauty, meaning or purpose – they would be meaningless if there were no such thing as experience

Why is Consciousness Important for AI ?

1) Is consciousness necessary for being intelligent ?

   Definitions of Intelligence, AGI (the ability to accomplish any goal at least as well as humans) and SuperIntelligence (intellect greatly exceeds cognitive performance of humans) **do not require consciousness to be present**

   Doesn't necessarily mean consciousness is not needed for intelligence.

2) Typical definitions of consciousness do not exclude possibility of artificial consciousness. **If future AI is conscious they may experience pain/suffering and so we should treat them accordingly !**

3) If future AI conscious to more or less same extent as humans, this may have profound cultural, social, political (as well as ethical) implications, dramatically changing the nature of human society. **What if we humans treat future AI as if they were conscious (irrespective y of whether they actually are conscious). What impact might that have on us ?**

We will return to 2) and 3) in the last lecture of the module

What are the Key Explanatory Problems ?

1) *The "easy" problem*

How does the brain work in all its functions, in all its detail. How does it interpret and respond to sensory inputs, how can it report on its internal state using language, how does the brain supports all behaviors ?

These are extremely difficult problems to solve, but they are problems of intelligence.

They don't require solving problem of what it **feels like/subjective experienc**e

2) *The "hard" problem*

How does physical matter give rise to feelings/experience – the *what it seems like from the inside*.

When you are driving you experience colors, sounds emotions and a feeling of self. Does self driving car experience anything at all ?

In both cases you are inputting information from sensors, processing it, and outputting motor commands. But subjectively experiencing something is *logically* separate – is it optional, and if so what causes it ? Why does one arrangement of particles result in experience and not another ?

Theories of Consciousness

In philosophy there are broadly speaking, two kinds of theories of consciousness

Dualist theories

- *Substance dualism* : both physical and non-physical (conscious) substances exist (Decartes 1644 “I think therefore I am”)
- *Property dualism* : two distinct and disjoint kinds of properties – physical ones and conscious ones.

Physicalist theories : everything is physical and property of consciousness can in principle be explained in terms of the physical

e.g. Functionalism which states that states/ processes count as being of a given mental or conscious type because of their functional role in a suitably organised system

- Also ***neutral monism***

A Philosophical Thought Experiment to show that Physicalism is false and Property Dualism is True

- **A philosophical zombie** is an exact physical duplicate of a person that behaves in exactly the same way as a person – you for instance – but who lives in a possible world that is indistinguishable from the actual world **except that such a zombie does not have experiential consciousness.**

Can you conceive of (imagine) such a possible “zombie” world ?

Philosophical Zombies

Consider the following argument:

- Zombies are conceivable
- If conceivable then possible - all the same physical properties of the actual world but not the experiential property of consciousness.

  Note that *conceivability* $\longrightarrow$ *possibility* is controversial
- Now if consciousness were a kind of physical property (X) it would be impossible for a creature to have the same physical properties as you but not have conscious experience (Y). But such a creature is possible ($\neg$ Y). Therefore consciousness is not a kind of physical property ($\neg$ X)
- Therefore *property dualism* is true.

Property dualism states that conscious properties are neither identical with nor explainable in terms of (i.e., reducible) to physical properties

Philosophical Zombies

To put it another way:

if everything that exists is physical or depends on (is explainable by) physical properties alone, then a world that is an exact physical duplicate will be a duplicate **in all respects**. So if there is a physical duplicate that is different in any way (doesn't have experiential consciousness) then *physicalism* is false.

But is such a zombie world even *conceivable* and so logically, let alone nomically possible ?

The more we understand about consciousness and the brain processes underlying experiential consciousness, the less conceivable such a zombie is …

- You may be able to imagine a 747 flying backwards, but the more you know about 747s, their structure, aerodynamics etc the harder it is to imagine/conceive.

Can you really conceive/imagine a philosophical zombie ?

Philosophical Zombies

But even if you reject the zombie argument and think that it is not conceivable, there is still an *explanatory gap* – the *hard problem*

How can we explain subjective feelings in terms of relations between material parts ?

The hard Problem Remains

- Everything you want to say about life - not conscious life - can be defined in terms of relationships amongst material parts, from metabolism to reproduction, even perception (light energy transformed to electrical and chemical energy and the mapping of the visual space to the visual cortex) - all can be explained in mechanistic terms/ultimately physics.

- But the exception is consciousness. In science there are explanations of how properties emerge from lower level physical facts Eg flow of water - molecules slide pass each other explaining higher level wetness - etc.

- But consciousness ? There are brute facts about how different conscious experiences correlate with brain processes – but these do not provide bridges between the physical and the feeling of what it is to be like …

- The hard problem still remains - how to account for the existence of such properties for which current science has no explanation.

Key issues in philosophical studies of consciousness and how

they relate to the possibility of conscious machines

What is a Conscious State ?

The notion of a conscious state has some distinct but interrelated meanings:

1. Simply a mental state one is aware of being in – to have a conscious desire for a coffee is not only to have such a desire but at the same time *to be aware you have the desire*
   (**in principle an AI could have such states of meta-awareness**)
2. Involves qualitative/experiential properties called *qualia*: raw sensory feels
   e.g., the taste of a glass of wine, the tactile feel of a piece of suede …

   Qualia are *phenomenal* properties. The *phenomenal* structure of *consciousness* refers to the overall structure of experience – not just sensory qualia but also much of the spatial, temporal and conceptual organization of our experience of the world and of ourselves as agents in it.
   (**could an AI have qualia ? The hard problem again**).

What is a Conscious State ?

3. A state might be conscious in the sense that it relates to, is available to, and interacts with other states. This is called ***access consciousness***

   E.g., a visual state's being conscious is not about whether there are associated qualia – the qualitative “what it's likeness” - but whether or not it and the visual information that it carries is available for use and guidance by the organism.

   (**in principle an AI could have such states**)

The Problems of Consciousness

1. **What is consciousness? What are its features ?**

- How does one give a description of qualia and the structure of phenomenological experience more generally ?

Is it indeed possible to give a *fully* descriptive account of phenomenological experience from the outside ?

Could there be *inverted qualia* – how do I know that what you experience when you see red is the same as I what I experience when I see red ?

The Problems of Consciousness

2. How can consciousness exist ?

Back to the ***hard problem*** –

Can we explain or understand how non-conscious items could cause/give rise to consciousness ?

Normally one analyses macro-properties in terms of their functions and then show how the micro-structures obeying physical laws are enough to guarantee satisfaction of the relevant functional conditions

The micro-properties of collections of H2O molecules at 20°C suffice to satisfy the conditions for the liquidity of the water they compose.

Moreover, the model *makes intelligible how the liquidity is produced by the micro-properties*.

However we have no such satisfactory explanation of how consciousness is produced ! There is an *explanatory gap.*

Some argue that given human cognitive limits we will never be able to bridge the gap, just as facts about democracy are *cognitively closed* to pigs

Some argue that the gap cannot *in principle* be closed, by any cognitive agents, and argue that this then means that physicalism cannot be true

If there is no way in which consciousness could be intelligibly explained as arising from the physical, it is not a big step to conclude that it in fact does not do so – that dualism is true – and ***if dualism is true this suggests conscious AI may not be possible***

The Problems of Consciousness

3. Why does consciousness exist ?

3.1. Can consciousness in principle play a causal role in behaviour ?

If our behaviours are ultimately determined by physical processes and our conscious experiences are non-physical, then how can conscious experiences play a role in causal chain of neuronal/chemical events that result in behaviour ?

Some think that consciousness is *epiphenomenal* - it doesn't play any causal/functional role (like steam from a train engine)

Indeed some features of consciousness – self awareness and awareness of one's own mental states such as when making a decision – may simply be ways of reporting to ourselves behaviours that are already underway

The Problems of Consciousness

3.2 If it does play a causal role, what is it ?

What is its adaptive value from an evolutionary point of view ? Note that its present function if it indeed it has one, might not be the same as the function for which it evolved.

Brains most basic function is to keep themselves and body alive – *interoception* – the feelings one is aware of in one's own body, emotions, mood, hunger, selfhood are arguably the brain's way of monitoring the body for signals indicating the need to act to ensure survival

Some more "complex" emotions/feelings evolved for social reasons, e.g., moral emotions such as disgust, shame, etc (social brain hypothesis)

Roles that Consciousness Plays

Supposing *epiphenomenal*ism is false, what sort of effects do features of consciousness have and what difference does it make ?

Important to understand if consciousness enables intelligent behaviour and whether machine might similarly benefit from consciousness

1. Self awareness and Awareness of Mental States/Meta-awareness

increases flexibility and sophistication of control (as compared with unconscious processes that by contrast are more efficient/speedy) which is especially important when dealing with novel situations and previously un-encountered problems – a hallmark of intelligence !

Roles that Consciousness Plays

Self awareness/awareness of one's own mental states also enables **social coordination**

Being aware of our own beliefs, goals, desires, perceptions enables us to have a better understanding of others' mental states

Mutually shared knowledge of others' minds (theory of mind) enables interaction, cooperation and communication in more advanced and adaptive ways - may be vital if we are to interact with the AIs of the future – reasoning together and acting together

Arguably such awareness of one's states – understood as having information about informational states - could be implemented in a machine

*(**without concerning ourselves with phenomenological character of "meta" mental states !**)*

Roles that Consciousness Plays

2. Conscious experience presents us not with isolated properties or features of reality, but **more unified and densely integrated representations of reality**:

Perception of objects independently existing in space and time relies on *integration* of info from various sensory channels as well as background knowledge & memory.

Not all sensory information needs be *experienced* in order to affect behaviour – eg. direct and reflex reactions in higher organisms and *robots*

But when experience is present, it provides a more unified and integrated representation of reality - one that typically allows for more sophisticated and flexible responses

But phenomenal experience may not be **necessary** for more unified integrated representations

Roles that Consciousness Plays

3. More global informational access. The information carried in conscious mental states is typically available for use by diverse mental subsystems and for application to a wide range of potential situations and actions.

Making information conscious typically widens its sphere of influence and the range of ways it can be used to guide or shape inner or outer behaviour

Information Theoretic Theories of Consciousness

Many notions of conscious states and the roles of consciousness (at least in terms of facilitating intelligent behaviour) have been described in informational terms: information about information (meta-awareness), integration of information, making information widely available ….

This suggests "purely informational" theories of consciousness that do not necessarily concern themselves with solving the "Hard Problem" of phenomenological experience

Indeed recent theories characterise or "explain" consciousness in terms of the storage, processing, integration and availability of information – ***information theoretic theories*** – and which therefore suggest that machine consciousness is possible

Global Workspace Theory

Global Workspace theory (GWT) describes consciousness in terms of a competition among processors and their inputs (sensory data, ideas …) and outputs "broadcasting" information for widespread access and use.

The particular processed outputs that "win out" over other processed outputs then become available to the *global workspace* and so accessible and influential with respect to other contents and other processors. At the same time recurrent support (feedback) back from the Workspace and from these other contents influenced by what is Broadcasted, "strengthens" the broadcasted content.

The availability and recurrent strengthening of the broadcast information makes it conscious in the **access consciousness** sense.

Integrated Information Theory

Integrated Information theory (*IIT*) originated by Giulio Tononi is also an information theoretic theory

Consciousness is linked to *the degree to which information is integrated*, which in turn can be represented by precise mathematical quantity Φ (phi)

Part of human brain supporting consciousness has very high Φ and is therefore highly conscious, whereas systems with a low Φ have a small amount of consciousness

So in principle we could have a consciousness meter that measures Φ and so tells us the extent to which another organism, or AI, is conscious

IIT implies ***panpsychism*** – a variety of (property) dualism that says all the constituents of reality have conscious, or at least proto-conscious, properties distinct from whatever physical properties they may have !

IIT's Criteria for Consciousness

A System having Information about itself

Books, photos, hard drives typically contain a lot of information about other things, which in turn depends on conventions about symbols and their meaning

According to IIT the information that matters for consciousness is the information the system has about itself – information about the systems *causal powers* :
how much can we know about the previous and next state of the system by looking at the state of the system right now?

Current state of typical human brain can tell you a lot about what brain looked like a moment ago, and what it will look like in the next moment – of course brain influenced by external conditions/environment but a lot determined by brain itself
(contrast with say human retina in which we learn little about past and future states as state almost completely determined by external conditions).

Integrated Information Theory (IIT)

Integration of Information

IIT's next requirement for consciousness – **integration** - measures (Φ) how much the information of a system depends on the interconnections between the system's parts. To determine it we ask: how much information is lost by cutting the system in two ?

Consider that If we tear the page of a book in half, almost no information is lost. Reading one half page and then the other conveys the same info as reading the intact page - the information on a page is not integrated and is reducible to the sum of the information of the parts.

But if highly interconnected brain cut in half much of what we can "read of" about prior and subsequent states will be lost - the information is highly integrated and not reducible to the sum of the information of its parts.

key difference **between brains and computers** is that although computers can have as much info as a brain – similar number of possible states - and have information about themselves, ***today's computers*** consist of transistors connected to only a few others, and that are feed-forward and not recurrent, and so have very low Φ

Integrated Information Theory (IIT)

Maximality of Integration

Third requirement is that a conscious system must be a *maximum* of Integrated information: it must have more integrated information than any of its parts and any bigger system of which it itself is a part.

Implications of IIT for AI

IIT rejects functionalist account of consciousness:

If a system S1 produces the same input/ output pairs as a system S2 that we know to be conscious, then SI is conscious, and whether S2 is made from the same kind of stuff (i.e., has certain properties such as high information integration) as S1 doesn't matter.

Why ? Because computers have sparse internal interconnectivity - consist of transistors connected to only a few others, and that are feed-forward and not recurrent, and so have very low integration (Φ)

Implications of IIT for AI

But in "Ascribing Consciousness to Artificial Intelligence" April 2015, arXiv Murray Shanahan asks us to imagine what would happen if we create a digital simulation of a brain in which we gradually replaced all the neurons with functionally equivalent digital equivalents and simulated all the connections –

Would this digital simulation really gradually lose consciousness ?

Also, nothing in principle prevents computers in the future from being built with sufficient integration

Arguably, IIT supports the idea that consciousness is the way **information feels** when being processed in *certain complex ways*

Implications for AI

If we take this idea seriously and reject IIT's insistence that digital computers cannot be conscious (by accepting the possibility that computers in the future may have sufficient integration), then arguably consciousness is *substrate independent*

So the information theoretic view of consciousness, together with substrate independence, suggest AI could be conscious

Conscious Machines

How would we recognise that an AI is conscious

- If it acts/behaves like it is conscious (conscious according to functionalism)
- High Φ or other information theoretic measures
- The right kind of architecture

Of course we could never, even in principle, *definitively* say that a machine is conscious since we cannot in principle experience the world from the inside in the way machine would, just as we cannot in principle experience the world from the inside in the way a bat or a dog or another human does

Also, interoception (monitoring internal state via conscious feeling) might have provided an evolutionary impetus for the origins of consciousness in humans.

What would account for the origins of machine consciousness ? Would we human designers have to intentionally design consciousness, or would it simply emerge as machines become intelligent, more sophisticated, as they process information In more and more complex ways ? ……