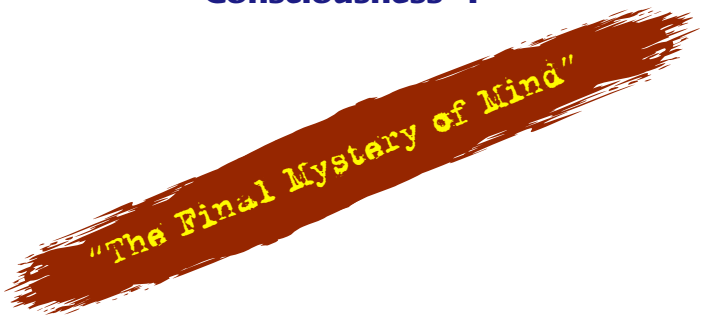
Part IV — Open Issues

Consciousness · I



Nov 23, 2017

A Slide from the second day of class …

1. Cognitive Science has mostly focused on B.

A) Subjective
1. Consciousness
2. Self-consciousness
3. Subjectivity
4. Feelings & sensations
5. Experience

B) Intentional / Semantic
1. Language
2. Thinking
3. Perception & action
4. Learning & memory
5. Curiosity

C) Normative/Affective
1. Emotions
2. Ethics
3. Spirituality
4. Religiosity (?)

D) Other
1. ~~Complexity~~

Primary historical focus of AI, cog sci, & phil-mind, & hence of this course

A Slide from the second day of class …

1. Cognitive Science has mostly focused on B.
2. Recently some attention has been shifting to A
3. This week we'll look mostly at A1, but also at issues of A2 and A3

A) Subjective
1. Consciousness
2. Self-consciousness
3. Subjectivity
4. Feelings & sensations
5. Experience

B) Intentional / Semantic
1. Language
2. Thinking
3. Perception & action
4. Learning & memory
5. Curiosity

C) Normative/Affective
1. Emotions
2. Ethics
3. Spirituality
4. Religiosity (?)

D) Other
1. Complexity

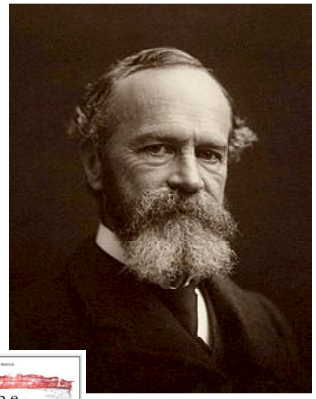
Back to Descartes (1596–1650) …

1. For Descartes: *mind ≈ consciousness ≈ self-consciousness*
   a) Consciousness, rationality, and conscious access to that rationality, were foundational assumptions
   b) "We cannot have any thought of which we are not aware at the very moment when it is in us"
   c) **Awareness** plays a huge role in the Meditations
   d) "(Self-)conscious rationality" was Descartes' *mark of the mental*

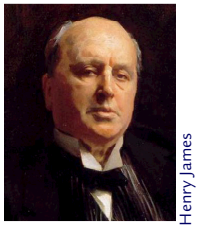
fast forward 250 years ...

William James (1842–1910)

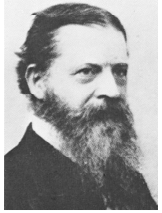
1. **Father of psychology**
   1. Brother: Henry James (1843–1916)
   2. Friend: Charles Saunders Pierce (1839–1914)
   3. Cf. Louis Menand's *The Metaphysical Club: A Story of Ideas in America*



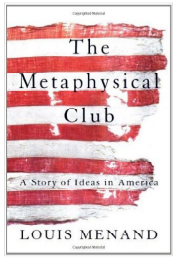
William James



Henry James

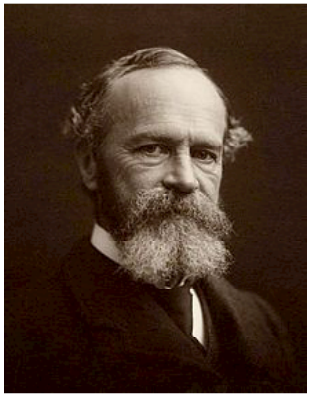


Charles Saunders Peirce



fast forward 250 years ...

William James (1842–1910) — cont'd



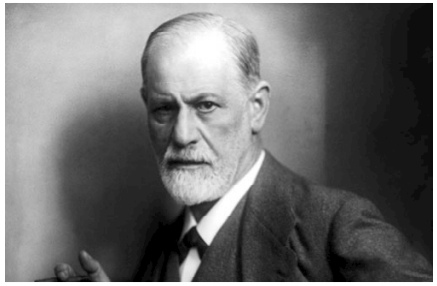
2. Started out thinking that consciousness was **essential to mind**
   - "Consciousness is the starting place of all psychology, the most crucial aspect of human mentality."
   - "Introspective observation is what we have to rely on first and foremost and always. Every one agrees that we there discover states of consciousness."

   — *Principles of Psychology*, 1890

3. Ended up thinking it should be **banished from scientific study**
   - "Consciousness is the name of a non-entity, and has no right place among first principles."

   — *Does Consciousness Exist?* 1904

Sigmund Freud (1856–1939)



1. One of the most important contributors to our contemporary understanding of consciousness ... through his (supposed) "**discovery of the unconsciousness**"
2. Striking contrast with Descartes—much that is true of us we are *not aware of* (something we now assume)
3. Developed over many years in (among other places) *psycho-analytic theory*

Why has cognitive science been so developed through an analysis of logic, rationality, etc.—instead of through an understanding of the psyche, including as delineated in psycho-analytic theory? ?

20th and 21st century

1. Early 20th c.: With rise of behaviourism and positivism (in part for political reasons!), consciousness banned from "proper" scientific discourse
   - Viewed as *subjective*, *epiphenomenal*, *unmeasurable*, *wholly inappropriate*
2. Continued (with notable exceptions) up through the 1970s
3. Suddenly, in the 1980s, consciousness went from **taboo** to **trendy**
   - Dennett: *Content and Consciousness* (1969) ⇐ *one of the exceptions*
   - Dennett: *Consciousness Explained* (1992)
   - Searle: *The Mystery of Consciousness* (1990)
   - Edelman: *Remembered Present: a Biological Theory of Consciousness* (1990)
4. Launch of the *Journal of Consciousness Studies* (1994)
5. Conference series: "Towards a Science of Consciousness"
   - 2014: April 21–26, Tucson, AZ
   - 2015: June 9–13, Helsinki, Finland
   - 2016: April 25-30, Tucson AZ
   - 2017: June 6-10, Shanghai, China
   - 2018: April 2–7, Tucson AZ ⇐ *you can go!*
6. Now a **huge** literature on consciousness in philosophy, neuroscience, ...

Polysemy — what does the term 'consciousness' refer to?

1. A global property that people (always) have—and that rocks & plants don't
2. A local property that you have while awake (and perhaps while sleeping?)
3. A property of beliefs, pains, etc. (e.g., "a conscious belief that Llewellyn is out to get him")
4. *Self*-consciousness ⇐ *how is it related to the others?*
5. "Pure" consciousness ⇐ *e.g., via meditation (chemicals?)*
6. Awareness or attention ⇐ *conscious of car outside house?*
7. Distinctive properties of the brain
8. … etc.

Polysemy — should we give them all distinct names?

1. A global property that people (always) have—and that rocks & plants don't ⇐ *Conscious1*
2. A local property that you have while awake (and perhaps while sleeping?) ⇐ *Conscious2*
3. A property of beliefs, pains, etc (e.g., "a conscious belief that Llewellyn is out to get him") ⇐ *Conscious3*
4. *Self*-consciousness ⇐ *Conscious4*
5. "Pure" consciousness ⇐ *Conscious5*
6. Awareness or attention ⇐ *Conscious6*
7. Distinctive properties of the brain ⇐ *Conscious7*
8. … etc. … ⇐ *Conscious8 …*

Isn't there something in common?

Issues

1. **General questions**
   a) What is it (that "thing in common")? ⇐ *in particular, and what kind of thing?*
   b) Is it amenable to scientific study?
   c) Does it supervene on physical materiality? ⇐ *can it be made out of straightforward physical stuff"?*
2. **Widespread agreement**
   a) Everyone knows *1st-person* (i.e., what it is like *subjectively*) ⇐ *i.e., in virtue of being conscious (cf. Louis Armstrong: "if you gotta ask, you ain't never gonna get to know)*
   b) No agreement *3rd-person* (i.e., on what it is like *objectively*)

Wide (if not wild) disagreement on whether it is amenable to scientific study

1. Some think **yes**
   a) Answer in neuroscience ⇐ *huge variety of different views*
   b) Answer in cognitive psychology
   c) Answer in quantum mechanics
   d) Answer will require reformulating science
      – Even: reformulating *what science is*
      – E.g., "pan-phenomenalism"
2. Some think **no** ⇐ *again, wide variety of views*
   a) Intrinsically unscientific ⇐ *so-called "new mysterians"*
   b) Not *intrinsically* unscientific, but we humans will never understand it…we're not smart enough!
   c) Someday, but not soon…

Four contrasting views

A. Subjectivity & indexicality
B. Inner awareness (introspection)
C. Coherence and control
D. Qualia

View A — Subjectivity and Indexicality

1. **First person**
   a) Private
   b) Authoritative
   c) Privileged access
   d) Not based on *evidence*, *perception*, or *sensation*
2. **NB: Doesn't require (Cartesian) *transparency***
3. **Perspectival**
   a) From a "point of view"
   b) Look out "from inside you"
4. **The "essential indexical"**
   a) Perry's "bag of sugar" story (from Tuesday)
   b) Indexicality in general: *me*, *you*, *today*, *here*, *there*, *yesterday*, etc....

As I said on Tuesday, my own view is that this derives from the differential equations of physics

"Of course that was long ago, but at the time it seemed like the present."

"Are we in this Starbucks or the one down the street?"

View B — Inner Awareness & Introspection

1. Something clearly right about this idea—or at least very common
2. "Truck-driver" phenomenon...
   a) Aware of the world
   b) Not aware of *being aware of the world*
3. Suggests the importance of **meta-level beliefs**
   a) Requires meta-level concepts (of 'belief,' 'pain,' etc.)
   b) Not otherwise perplexing?
4. In my view, a better story about **self-consciousness** than about ***consciousness per* se**
5. NB: Meta-level cognition cannot be the story of what secures the *ineliminable first-person character of consciousness* (cf. View A)
   a) One's meta-level beliefs must still recognize object-level beliefs *as one's own*
   b) Cf. Zahavi's *Self-Awareness & Alterity* (1999)

The subject matter of my PhD dissertation (on "reflection")

View C — Coherence & Control

1. Dubbed **access consciousness** by Ned Block
   a) Available for epistemic self-reporting
   b) Able to control behaviour
2. Unification of sensory information
   a) Multi-modal integration
   b) 40Hz brain waves (and subsequent versions)
3. Seems the most likely thing to be explained *at the neuroscientific level*
4. Status
   a) Unarguable that this kind of functionality *exists*—and is *crucial*
   b) Yet to many (including me!), it does not seem (enough) to explain (what are viewed as) those properties of consciousness that are most distinctive, that involve first-person subjectivity, and that warrant its being considered at least in some ways mysterious.

View D — Qualia

1. The "**what it is like to be**" (sensory?) character of conscious mental states, as characterized by Thomas Nagel (in "What Is It Like to Be a Bat?")
2. For many, this is where the "mystery" of consciousness resides
3. **Qualia**: The *raw feel*
   a) The "redness of red"
   b) The "nutmeg-y taste of nutmeg"
   c) ... etc.
4. Has to do with (is?) the *qualitative* or **phenomenological character of experience**
5. Leads to "what it is like" to be in a mental state
   a) Seems immediately applicable to *sensory* and *emotional* states
   b) Less clear (there is disagreement) over states of *belief*

Possibly the deepest characterization of consciousness

The "Hard Problem"

1. Famously, David Chalmers has dubbed the problem of explaining qualia the "**hard problem**"
2. Cf. discussions of **zombies** ⇐ *a "culturalist", inaccurate reference!*
   a) Putatively, zombies are "*just like us*" (mechanistically *and semantically*) but lack any qualia, any "qualitative sense of being them"
   b) Huge debates on whether the notion has any intellectual merit
   c) But a great deal of literature has been written exploring the idea
3. Arguments (*pro* and *con*)
   a) Inverted spectrum
   b) *Logical* vs. *nomological* vs. *metaphysical* vs. *empirical* (im)possibility
4. Cf. tooth pain example (Güzeldere)

From Qualia to the Explanatory Gap

1. "Explanatory Gap" formulation by Joe Levine
2. How do we deal with the (seemingly) vast, dreaded gap between
   a) **Subjective:** the *involved*, inexorably *first-person phenomenological* or *qualitative* character of *conscious experience*, and
   b) **Objective:** the *detached, third-person* character of empirical science—from physics and neuroscience to cognitive science to scientific psychology?
3. Can these two views be unified?

The best formulation of the challenge facing cognitive science (imho)

No problem, say I!

Next Tuesday I will talk about how I think qualia arise, and about how the explanatory gap can be crossed

Consciousness · II

Nov 28, 2017

Getting underneath the objects ...

1. The reason that the "hard problem" (qualia) is so hard, I believe, has to do with a fundamental assumption that has not only been made throughout the history of cognitive science, but that also permeates a great deal of analytic philosophy.
   a) It also has to do as well with "failure" of GOFAI
   b) And with the challenges of the "frame problem" (and issues of "relevance")
   c) And with various other missteps in the history of cognitive science
2. The problem is people's—but especially theorists'—**ontological assumptions**
   a) In this respect I agree with Bert Dreyfus (in his 4th philosophical critique)
3. In particular: to **assume** (without question) that the world consists of *objects*, *properties*, *relations*, etc. (i.e., to assume, as I would put it, that the world comes **pre-registered**) blocks our ability to understand all four of the following:
   a) **What consciousness is**
   b) **How consciousness works**
   c) **How it arises from the fundamental nature of physics**, and
   d) **Why consciousness has the distinctive phenomenological feel**—*qualitative character*—that it does.

Getting underneath the objects (cont'd) ...

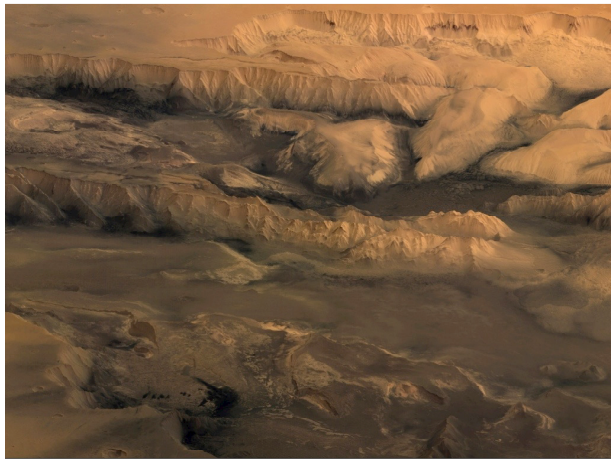
4. Assuming that the world consists of unproblematic *objects*, *properties*, *relations*, etc also blocks our ability to understand ...
   a) The role of *perception*
   b) The importance of *embodiment* and *(en)action*
   c) The (huge) merits and (substantial) demerits of **inference** and **conceptual reasoning**
5. In fact, to put it starkly:

Assuming that the ontology of the world is given blocks our ability to understand the mind

Can we understand the world without objects?

1. Yes, I think we do *all the time!*
2. We have already seen a bit what that might be like
3. I.e., what the nature of the world might be, below the level at which it is "parsed" into recognizable objects, properties, relations, etc....

For example—and importantly—I do not believe that the *natural environment* can be accounted for in terms of any such discrete ontology...



Similarly...

And ...

And ...

And ...

And ...

1. The question is how cognitive science would be different if we
   a) Didn't start assuming that the world consisted of discrete objects, properties, relations, etc., and
   b) Instead assumed it was structured in such a way as to make these natural environments more easily intelligible?
2. I will argue that doing so means **adding a third intellectual challenge** to the problem of explaining the mind.

Three major intellectual challenges for Cognitive Science

#1 — Naturalizing Semantics and Intentionality

Well recognized

1. It is generally recognized, in philosophy, that if we are to have an understanding of the mind that is either itself scientific, or (perhaps more importantly) meshes seamlessly with science, we have to have accounts of mental phenomena, or concepts that are used to describe mental phenomena, that show *how they can arise from*, or *how they are ontologically compatible with*, the world as described in the natural sciences.
2. This theoretical project is called **naturalizing** such concepts or notions or phenomena.
3. The most famous example is **naturalizing intentionality** or **naturalizing semantics**
   a) "Naturalizing intentionality" is a project that any philosopher will recognize
4. The basic question is *how semantic relations* (the ones we have indicated all semester with blue arrows—like reference to distal objects in the world) can be explained "in scientific terms". (⟹)
   a) One reason this is challenging, as we have seen, is that semantic relations of reference *aren't (directly) causal* in any evident sense.
   b) In general, it is not at all clear how *meaning something* (such as that dinosaurs were warm-blooded), or *being true*, or *referring to the Pharaohs of Egypt*, can be accounted for in scientific terms.

Three major intellectual challenges for Cognitive Science (cont'd)

#2 — Naturalizing Normativity — "Ought" from "Is"

Also well recognized

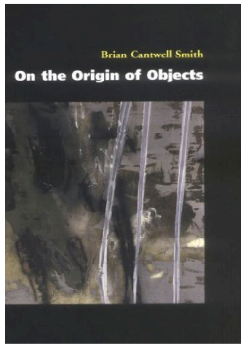
1. Another huge—and famous—issue in philosophy has to do with **naturalizing normativity**
2. I.e., understanding how to make "scientific" sense of **normative** or **evaluative** notions—such as some things being:
   a) *Good*
   b) *Better* than other things
   c) *Just*
   d) *Beautiful*
   e) *Worthwhile*
   f) … etc.
3. Similarly: such ultimately normative notions as being *kind, altruistic, helpful,* etc.
4. The question is often put as one of whether we can derive "ought" from "is"
5. Famously, some people think this cannot be done
6. Recently, a huge movement has arisen that attempts to naturalize normativity by deriving what is "good" from "what is **evolutionarily advantageous**" — though I myself am not a fan of that approach.

Three major intellectual challenges for Cognitive Science (cont'd)

#3 — Naturalizing *Ontology* — Objects, Properties, Relations, etc.

New! Needs to be added!

1. The issue that I think we have to take on, in cognitive science, if we are to understand consciousness (and a whole bunch of other things!) is that we have to:
   a) **Naturalize ontology!**
2. That is, we need to understand—as I put it—how people **register the world**—find it intelligible, take it to consist of objects, properties, relations, etc., or to consist of any other sort of thing!
3. Thus I would say:
   b) I register *a car* (while looking out the window)
   c) She registered him *as a threat*
   d) The two of them were speaking *in different registers*
   e) The way in which the elder registers the situation is almost entirely incomprehensible to me
4. I describe this theory of registration is my book *On the Origin of Objects*



1. We don't have time to talk about registration in general today
2. What I want to do is to show how, at the lowest level, it gives us a glimpse of what could underlie the qualitative character of experience (qualia)
3. So start at the beginning (or at least at what people who think there is a beginning think you can start with!)—i.e., ***fundamental physics***

Question #1 — What is the world like, according to physics, if we don't assume objects?

1. The best way to think of this is field-theoretically
2. It is a world of spectacular and stunning complexity
3. A stupefyingly complex superimposition of interpenetrating waves, vortices and fields and quiescence and turbulence—vibrations from glacially slow to blazingly fast, forces continuously impinging, forces welling up and falling continuously away
4. Imagine falling overboard in a storm at sea, surrounded by nothing but crashing waves, stinging spray, and undulating currents, as far as the eye can see

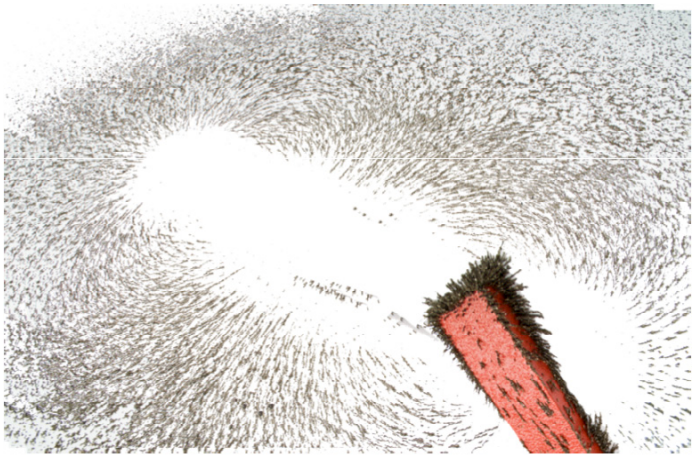
   —***and then subtract you!***

5. That is approximately what the world is like, according to physics—except a zillion times worse.

Question #2 — How does physics *work*, if we don't assume objects?

At any given place and time in the 4D physical plenum, there are **point-to-point interactions** with neighbouring places and times—i.e., with those space-time points that are **spatially and temporally adjacent**

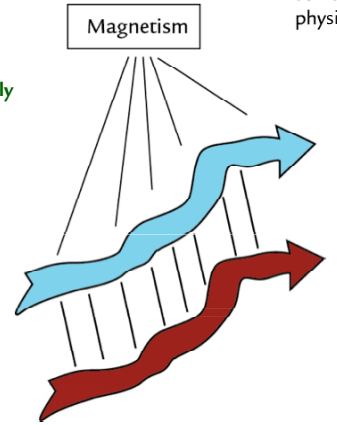
Example



Example

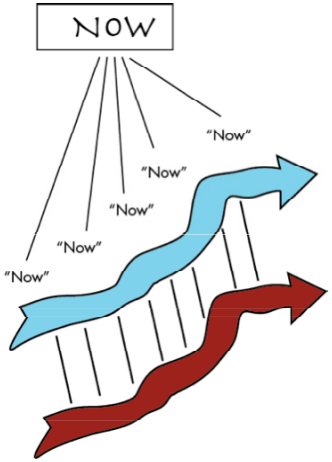
The **point-to-point correspondence** of physical laws

This point-to-point correspondence (of all physical regularities) is **strikingly similar to the way that indexical or deictic references work in natural language** (*here*, *now*, *I*, *today*, etc.)



Deixis / indexicality

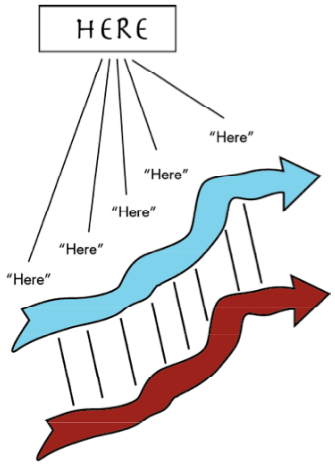
This point-to-point correspondence (of all physical regularities) is **strikingly similar to the way that indexical or deictic references work in natural language** (*here*, *now*, *I*, *today*, etc.)



Deixis / indexicality (cont'd)

This point-to-point correspondence (of all physical regularities) is **strikingly similar to the way that indexical or deictic references work in natural language** (*here*, *now*, *I*, *today*, etc.)

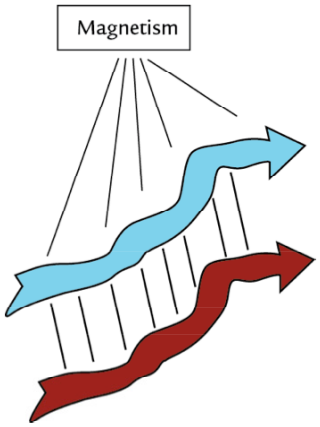
the crucial point!



Deixis / indexicality (cont'd)

It is as the magnet were constantly talking to the iron filings:

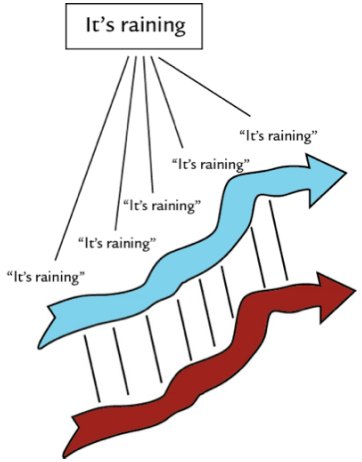
"You! Come here! Now!"



Deixis / indexicality (cont'd)

Deixis underlies a *huge amount* of language—not just obvious indexicals, but other common forms that don't posit or require (i.e., that don't register the world in terms of) *discrete individual objects*.

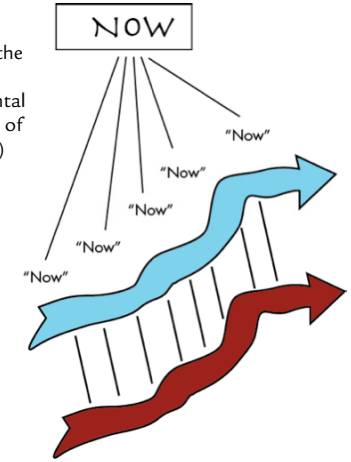
Cf. what Strawson calls "feature placing"



Deixis / indexicality (cont'd)

Because *physical (causal) interaction* has the same structure as *deictic or indexical language*, I say that there is a fundamental **deixis** (indexical structure) to the laws of physics (i.e, to all physical regularities)

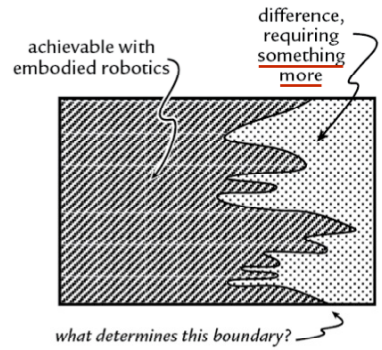
the ontological consequence of the epistemological fact that the laws of physics are expressed as differential equations



How does this all relate to *objects* … and to *representation*?

1. I believe they are profoundly related
2. As we said, I believe that cognition involves a process of *registering* the world in terms of objects, properties, etc., and that
3. Registration is a process of **representing the world as** consisting of objects, properties, features, etc.…
4. How does that object-registration go?
5. Go back to Brooks and dynamic robots, and the "something more" that being in the world required

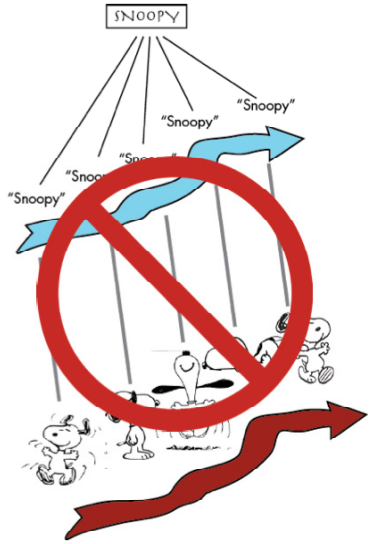
How does this all relate to *objects* … and to *representation*? (cont'd)



slide from lecture 10c

… and that "more" is ***representation***
(normatively governed non-effective relations to what is distal)

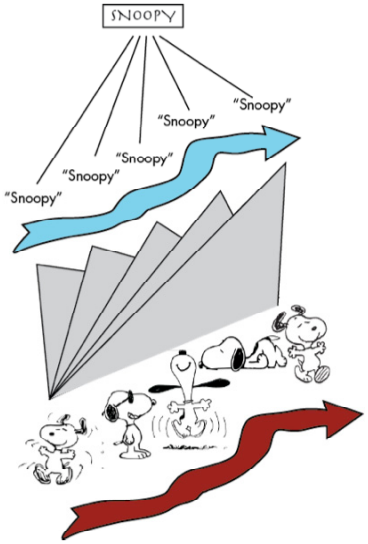
Object registration is *not simply a case of point-to-point correspondence*



✗

Rather, object registration involves *point-to-extent* correspondence

1. *Each* utterance of a name refers to the *entire temporally-extended object* that is its reference
2. It opens up the whole (huge!) metaphysical subject matter of the relations between ***the one and the many***

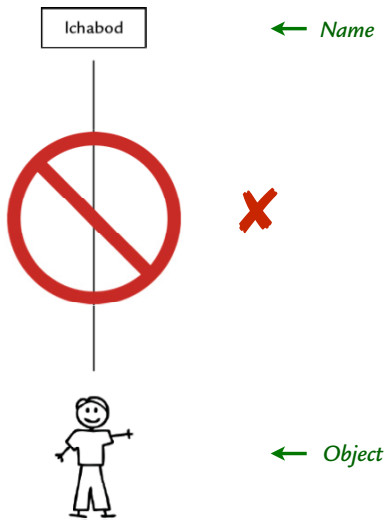


✔

In other words, the

one name ↔ one object

model that children learn in school is far too simple.



In other words, the

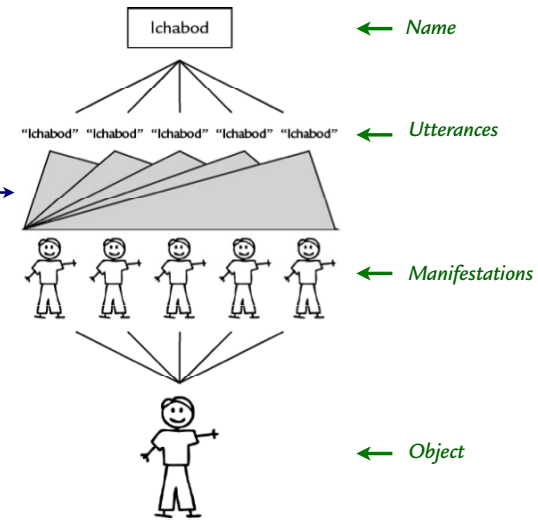
one name ↔ one object

model that children learn in school is far too simple.

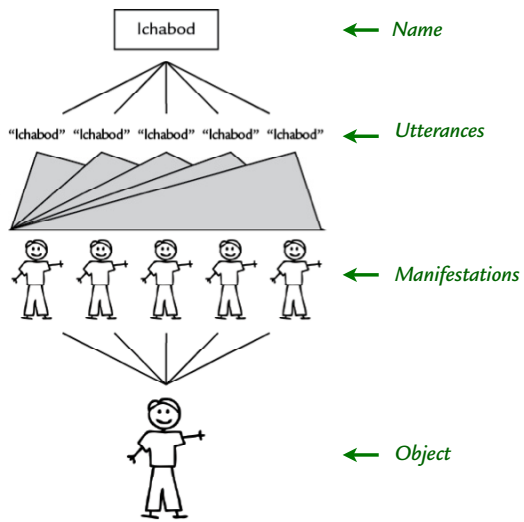
It is more like

one name ↔
many utterances ↔
many manifestations ↔
one object

where—crucially!—
each utterance refers
to *all* manifestations



This *point-to-extent* nature of object representation (registration) is one reason why objects can only arise via the ***disconnection*** (non-causal coupling) that is true of representation in general (as we have seen since the beginning)!

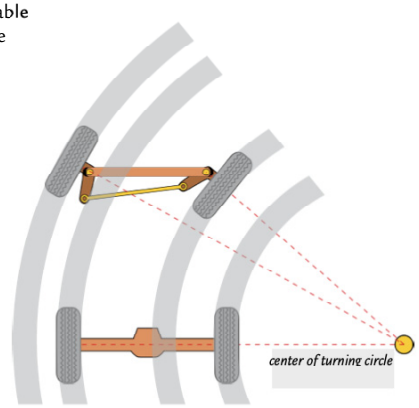


That is ...

The **object of our consciousness** is *never something we are directly causally connected to!*

Rather, what we are conscious of is a stable entity (point, phenomenon, object, state of affairs, etc.) to which we maintain a (disconnected!) relationship

Analogy: steering the wheels on a car (to prevent skidding): the Watermann model

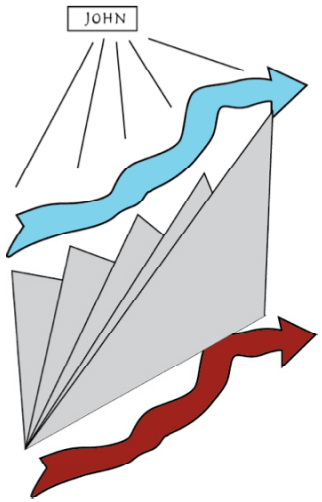


Stabilizing reference

Registering objects requires **stabilizing the world** (not stabilizing one's self)

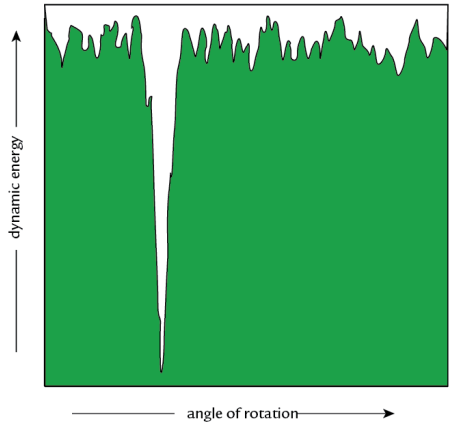
Stabilizing the world requires **compensating for change**

Compensating for change may require changing **oneself**, in order to **hold the world stable**



Example: the vestibulo-ocular reflex

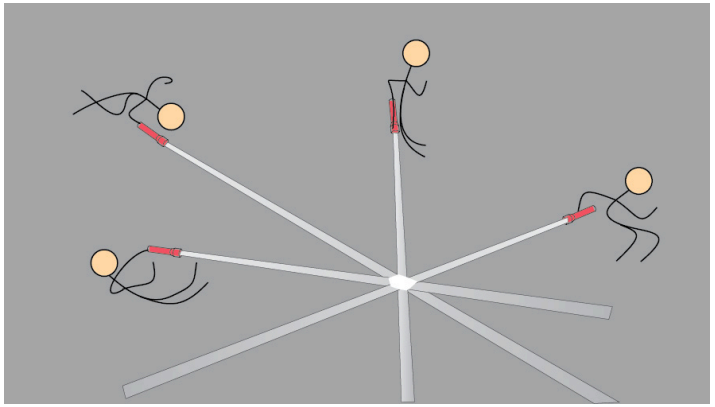
By moving one's body (or brain, or neurons) one can stabilize *that on which one is focused*



Stabilizing reference

"Behind me"	⟹	"In front of me"	(*upon turning around*)
"Yesterday"	⟹	"Today"	(*when you wake up tomorrow*)
"Me"	⟹	"You"	(*in our conversation*)
"The tallest person in the class"	⟹	"The second tallest person in the class"	(*when someone new arrives*)
<43° 40' 1.16" N, 79° 23' 30.7" W>	⟹	Here!	
Ebeneezer Le Page	⟹	S.I.N 876-543·210	
...	⟹	...	

The 'Intentional Acrobat'



Consciousness

Fluid, flexible, continuous, extraordinarily detailed practices of **deconvolving the** (underlying physical) **deixis**, in order to **stabilize the distal world**, consisting of a myriad forms of *indexicals*, *features*, and *objects*, intimately coupled with facilities for *movement*, *navigation*, and *survival*

If one takes the field-theoretic nature of the physical plenum seriously, and realizes what is involved in these registration practices, there is:

No mystery as to why phenomenal consciousness has the phenomenal feel that it does!

Ethics of AI

We are honoured to have Atoosa Kasirzadeh present today's lecture on one the most important issues facing society today.

Nov 30, 2017

Ethical considerations of Artificial Intelligence

Atoosa Kasirzadeh

University of Toronto

2017-11-30

Why ethical considerations of AI ?

ROBOSTOP Facebook shuts off AI experiment after two robots begin speaking in their OWN language only they can understand

Experts have called the incident exciting but also incredibly scary

By James Beal and Andy Jehring
1st August 2017, 12:03 am | Updated: 2nd August 2017, 4:56 am

39 COMMENTS

FACEBOOK shut down an artificial intelligence experiment after two robots began talking in a language only they understood.

The "chatbots" Alice and Bob modified English to make it easier for them to communicate — creating sentences that were gibberish to watching scientists.

Intelligent Machines

Military Robots: Armed, but How Dangerous?

The debate over using artificial intelligence to control lethal weapons in warfare is more complex than it seems.

by Will Knight August 3, 2015

Weapons are becoming increasingly automated.

22

An open letter calling for a ban on lethal weapons controlled by artificially intelligent machines was signed last week by thousands of scientists and technologists, reflecting growing concern that swift progress in artificial intelligence could be harnessed to make killing machines more efficient, and less accountable, both on the battlefield and off. But experts are more divided on the issue of robot killing machines than you might expect.

John Giannandrea.
GETTY

Intelligent Machines

Forget Killer Robots—Bias Is the Real AI Danger

Television
The other side

The Sex Robots Are Coming: seedy, sordid - but mainly just sad

The sex-doll industry is going from strength to strength in the drive to make figures more lifelike, but where will it end?

11,779 102

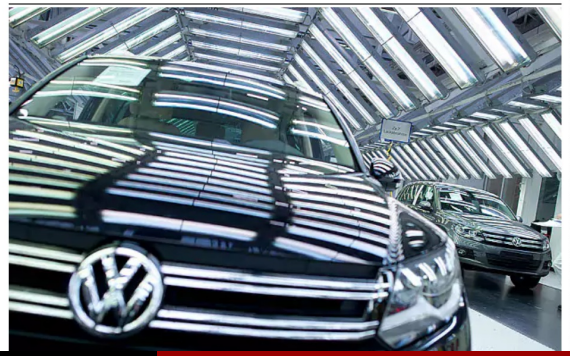
Fiona Sturges

Saturday 25 November 2017 11.00 GMT

HOME » NEWS » WORLD NEWS » EUROPE » GERMANY

Robot kills man at Volkswagen plant in Germany

A 22-year-old worker was grabbed by the robot and crushed against a metal plate



Varieties of ethical issues concerning AI

- What ethical principles should AI researchers follow ?
- Are there restrictions on the ethical use of AI ?
- What is the best way to design AI that aligns with human values ? (which values ?)
- Is it possible or desirable to build moral principles into AI systems ?

Varieties of ethical issues concerning AI

- Are AI systems themselves potential objects of moral concerns ?
- What moral framework and value system is best suited to assess the impact of AI ?
- Robots will come in and take some jobs (shift in the workforce)

Short-term issues

- Ethical considerations of driver-less cars
- Algorithmic discrimination
- Unmanned air vehicles (drones)
- Care-assistant robots
- Field being dominated by white male dudes

Long-term issues

- Threats to Humanity's survival
- Global catastrophic risk : a hypothetical future which has the potential to damage human well-being on a global scale

Moral philosophy in a nutshell

- Ethics is concerned with the study of values : what is good or bad and what is right or wrong
- Different ethical theories determine right from wrong by focusing on different principles

1. Descriptive ethics : behavior and thought of people when dealing with moral issues
2. Normative ethics : principles and theories guiding our moral actions
3. Meta-ethics : meaning and structure of moral beliefs and the origin of our ethics
4. Applied ethics : application of moral norms to specific moral issues (non-human animal rights ; nuclear wars)

Introduction to some ethical theories

1. Utilitarianism
2. Deontology
3. Virtue ethics

1- Utilitarianism

- Pioneering figures : Jeremy Bentham (1748-1832) and John Stuart Mill (1806-1873)
- The most ethical choice is the one that produces the greatest good for the greatest number
- Utilitarianism determines right from wrong by focusing on the **outcomes (consequences)**

Utilitarianism's challenges

- Whose utilities we should maximize ?
- How much we can be certain of consequences of actions in an uncertain world ?
- How should we define utility ?

2- Deontology

- Deontological theories focus on the **right action** (as opposed to consequences)
- Many deontologists believe that some actions are forbidden **no matter how good are the consequences**
- Main maxim : people should follow the moral rules and do their duties
- An instance of Deontological theory was proposed by Immanuel Kant (1724-1804)
- Kantian categorical imperative : act only on the maxim through which you consider the maxim to become a universal law

Challenges to deontology

- Too restrictive
- Who should determine what those rules are ?

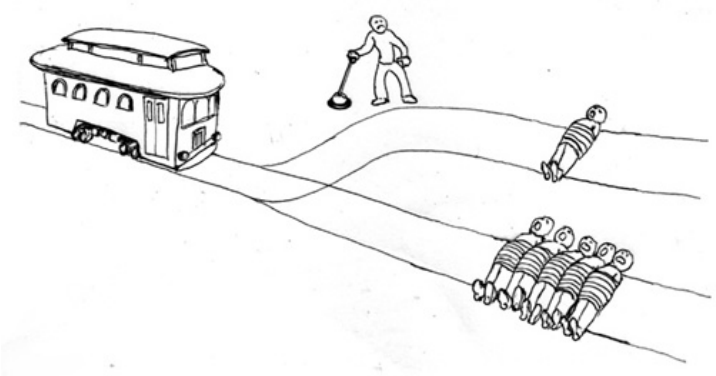
3- Virtue Ethics

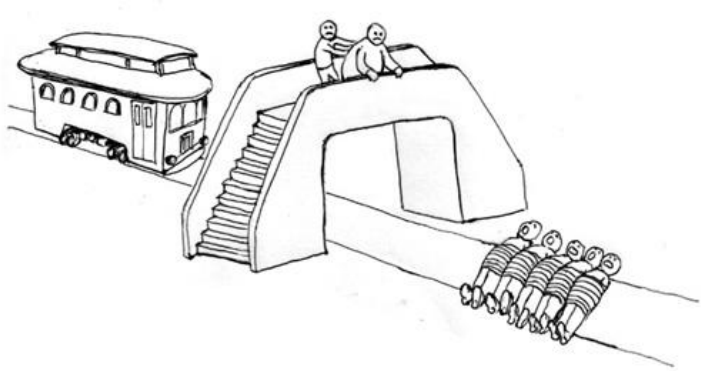
- Acquisition of virtues through practice
- Character- and person-based approach to morality : we acquire virtue through practice
- A quest to understand and live a good life of the moral character
- Introduced by Confucius, Aristotle, and other ancient philosophers

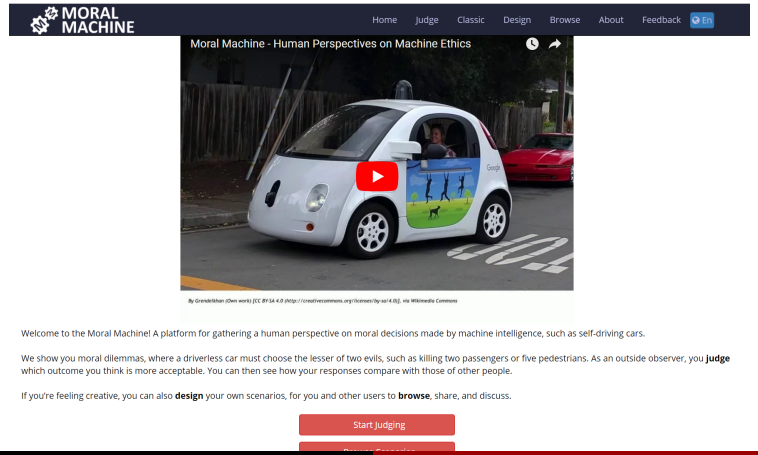
Virtue ethics' challenges

- It cannot provide responses to moral dilemmas
- It does not benefit the individual
- How should we agree on what are virtues ?

Trolley Problem
(a philosophical thought experiment proposed by Philippa Foot (1967))







Some challenges to keep in mind

- Should we use human moral theories or should we build a moral theory for machines ?
- Robot rights are like non-human animal rights ?
- How should we treat a robot that robs a bank ?
- Some research along this challenge : Sacrifice One For the Good of Many ? : People Apply Different Moral Norms to Human and Robot Agents (Malle, Scheutz, Arnold, Voiklis, Cusimano 2015)

How to start ?

- Formalism of ethical theories such that they lend themselves to algorithmic implementation
- Implementation of moral reasoning in autonomous systems
- Several problems about codification of ethical theories

The robots as an ethical agent

- Moor's classes of ethical agents :
  1. Ethical impact agents : their actions have ethical consequences (whether intended or not)
  2. Implicit ethical agents : ethical considerations built into their design (mainly safety or security considerations)
  3. Explicit ethical agents : identify and process moral information about different situations and make sensitive normative decisions about what should be done
  4. Full ethical agents : make moral judgments (and state justification about their judgments

Robot Ethics

Robot ethics

1. Ethical questions about the design, development, and deployment of robots
   - Ethical code of conduct and guidelines for engineers and policy makers
   - Restricted use of robots (e.g., autonomous weapons)
   - Potential ban on sex robots (criticism : robots reinforce objectification and exploitation of females)
2. Ethical questions about the users of robots

Robot ethics

- Kenji Urada (died 1981) was one of the first individuals killed by a robot. Urada was a 37-year old maintenance engineer at a Kawasaki plant. While working on a broken robot, he failed to turn it off completely, resulting in the robot pushing him into a grinding machine with its hydraulic arm. He died as a result.

Guidlines for ethical robots

- First attempt : three laws of robotics by Isac Asimov (for sci-fi) :
  1. A robot may not injure a human being or through inaction, allow a human being to come to harm
  2. A robot must obey orders given by human beings except where such orders would conflict with the first law
  3. A robot must protect its own existence as long as such protection does not conflict with the first or the second law

Machine Ethics

Learning morality ?

- Learning as improving performance over time
- Learning as a function which predicts output given a set of inputs-outputs
- Kinds of learning : supervised, unsupervised, semi-supervised, reinforcement learning

Engineering moral machines : top-down vs. bottom-up

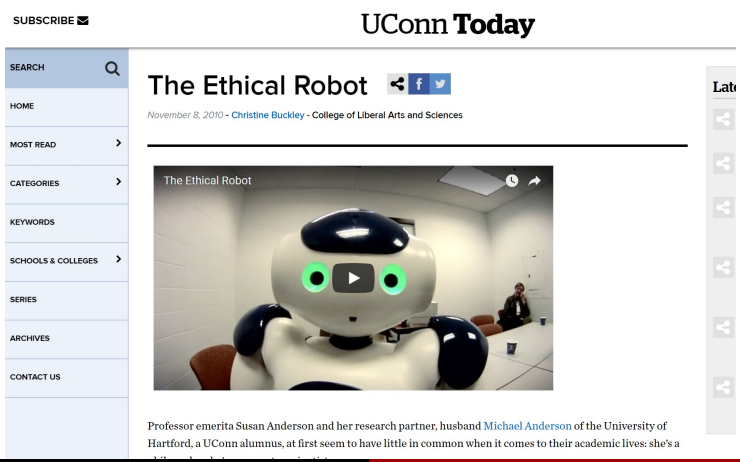
- Top-down strategies : implement (selected) normative theories of ethics and ensure that the moral agent acts aligned with the principles underlying the theory
- Bottom-up strategies : ethical theories emerge via the activity of individuals rather than in terms of normative theories of ethics

Supervised learning

- Given a training set of N example input-output pairs, in which the output elements are generated by an unknown function $g(x)$, find a function $f(x)$ (hypothesis) that approximates the true function $g(x)$
- Calculate the accuracy of $f(x)$ with a test set of inputs to which we know the right output
- When the output values are from a finite set, the learning problem is called classification

Reinforcement learning

- Learning from a series of rewards and punishments
- Abel, MacGlashan and Littman (2016) : the ethical learning and decision making
- Armstrong (2016) : the ethical decision making and learning as Bayesian learning problem

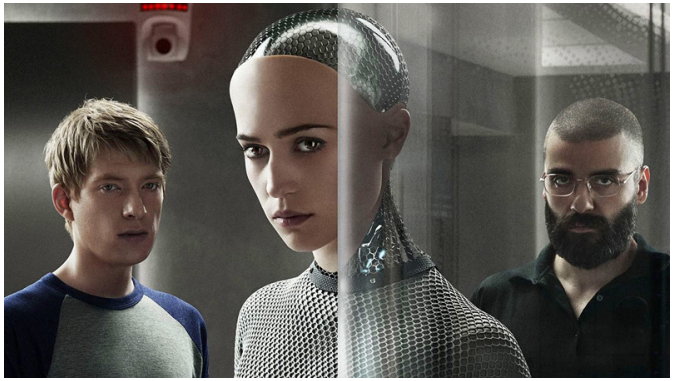


Ethical challenges of data science

Some research questions

- What is data ethics ? (Floridi, Taddeo 2016)
  - How data is generated, recorded and shared : the ethics of data
  - How AI, machine learning and robots interpret data : the ethics of algorithms
  - What ethical elements can be assigned to the corresponding practices (responsible innovation, programming, hacking and professional codes)

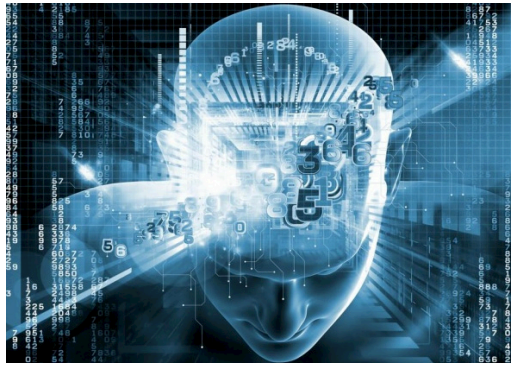
Thanks for your attention



Part IV — Open Issues

The Singularity

1. Characterization
2. Kurzweil
3. Possibility (Chalmers)
4. Merits & demerits
5. Reactions
6. Prospects ...



Singularity · 1 — Characterization

"Let an ultra-intelligent machine be defined as a machine that can surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design an even better machine; there would unquestionably be an intelligence explosion and the intelligence of man would be left far, far behind.

Thus, the first ultra-intelligent machine is the last invention that man need ever make."

— Irving J. Good, "Speculations Concerning The First Ultra-intelligent Machine," 1965

Sorry! ...

Singularity · 1 — Characterization (cont'd)

"What are the consequences of this event? When greater-than-normal intelligence drives progress, that progress will be much more rapid. In fact, there seems no reason why progress itself would not involve the creation of *still more intelligent entities*—on a still-shorter time scale.

The best analogy that I see is with the evolutionary past. Animals can adapt to problems and make inventions, but often no faster than natural selection can do its work—the world acts as its own simulator in the case of natural selection. We humans have the ability to internalize the world and conduct "What if's" in our heads; we can solve many problems **thousands of times faster than natural selection**. Now, by creating the means to execute those simulations at much higher speeds, we are entering a regime as radically different from our human past as we humans are from the lower animals.

From the human point of view, this change will be throwing away of all previous rules, perhaps, in the blink of an eye, an exponential runaway beyond any hope of control."

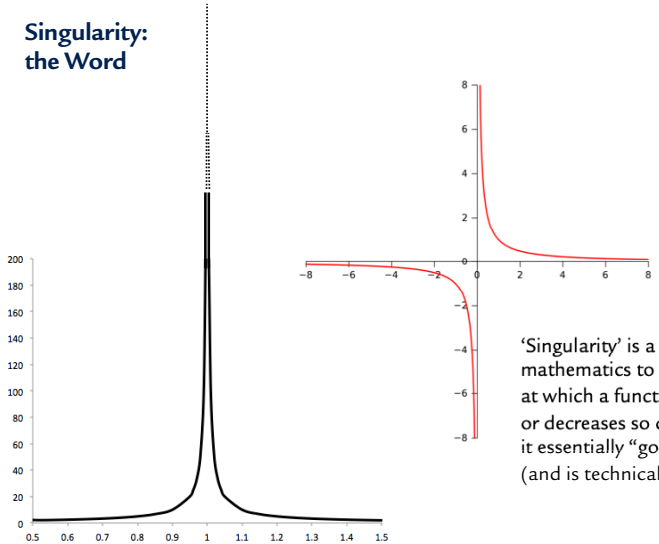
— Verne Vinge, 1993

The argument for the singularity

	Whether	When
1. Computational power will continue to grow exponentially	?	?
2. Intelligent machines will exist, via:		
a) AI succeeding,		
b) Brain simulation; or	?	
c) (Simulated?) evolution		
3. Computers will be more intelligent than us	?	?
4. Then, or soon thereafter, they will be capable of building *yet more powerful computers*	?	?
5. Machine intelligence will then increase explosively, leading to a **singularity**		?
6. All this will happen soon		
a) In your lifetime	?	
b) After you have died		

What do you think?

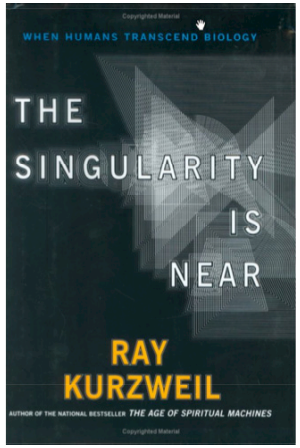
Singularity: the Word



'Singularity' is a term used in mathematics to signify a point at which a function increases or decreases so drastically that it essentially "goes to infinity" (and is technically undefined).

Singularity · 2 — Ray Kurzweil, Its Prophet

2005



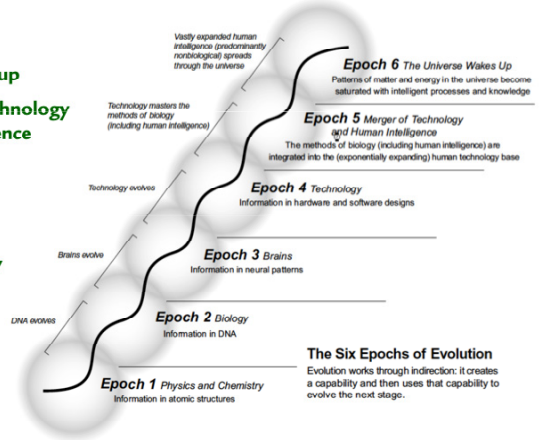
"First we build the tools, then they build us."
— Marshall McLuhan

"The future ain't what it used to be."
— Yogi Berra*

*Or perhaps Paul Valéry, or Laura Riding, or Robert Graves, or …

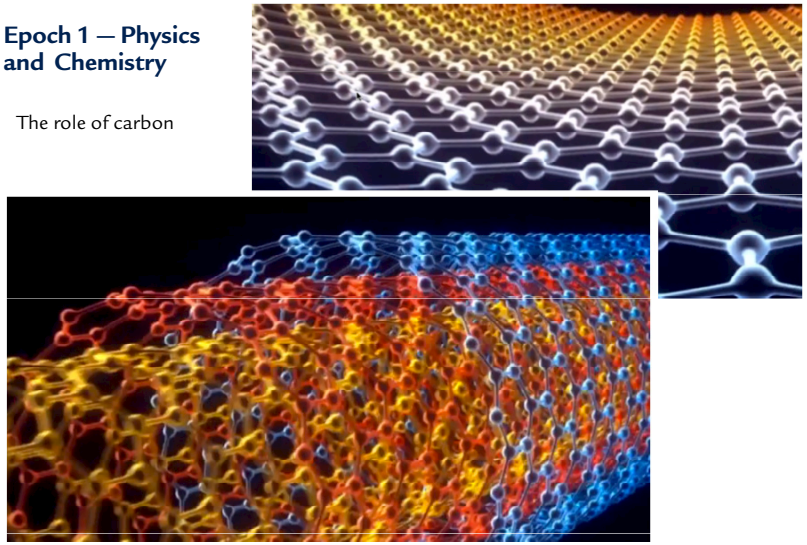
Kurzweil's "Six Epochs"

6. The universe wakes up
5. Merge of human technology with human intelligence
4. Technology
3. Brains
2. Biology and DNA
1. Physics & Chemistry



Epoch 1 — Physics and Chemistry

The role of carbon



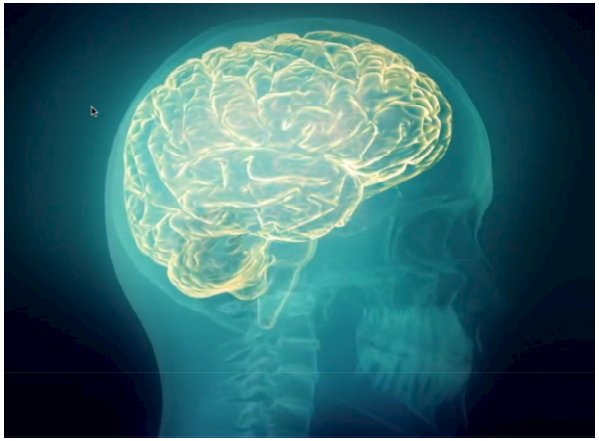
Epoch 2 — Biology & DNA

1. Information
2. Records (DNA) so that evolution could keep track of, and benefit from, its ongoing experiments

Epoch 3 — Brains

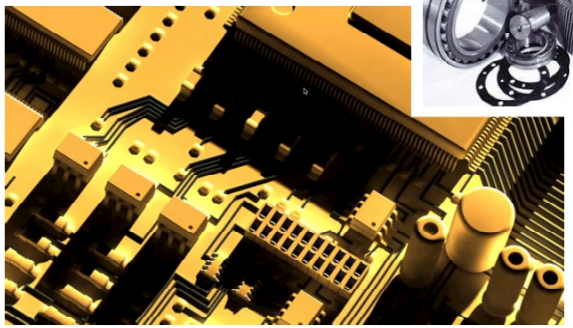
1. **Representation**!
2. **Mental models** of the world!

Does this sound familiar?

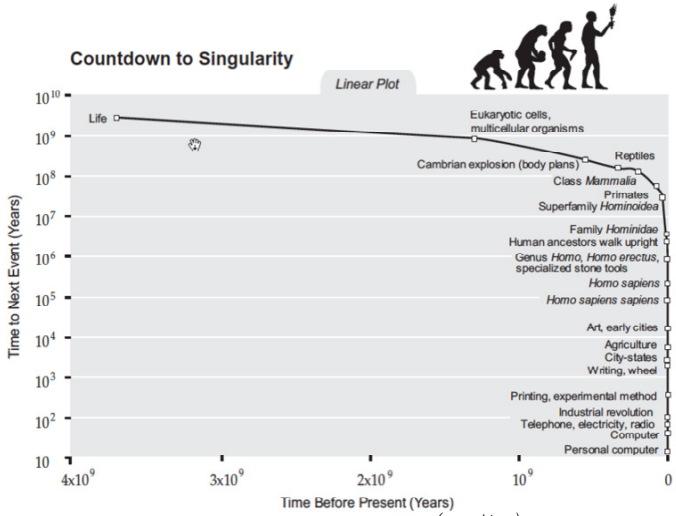


Epoch 4 — Technology

1. Mechanical tools
2. Information technology



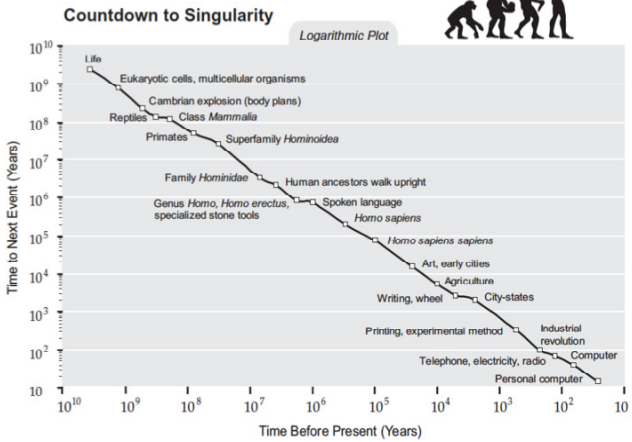
The Exponential Pace of "Progress"



Linear view of evolution: This version of the ~~preceding~~ figure uses the same

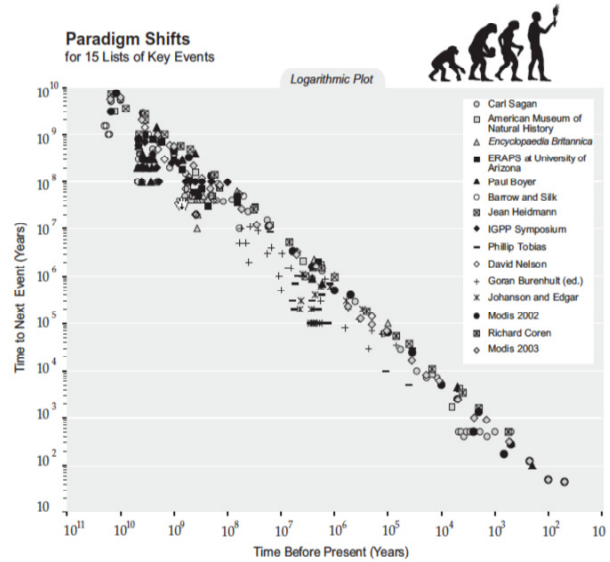
(see next page)

The Exponential Pace of "Progress" (cont'd)



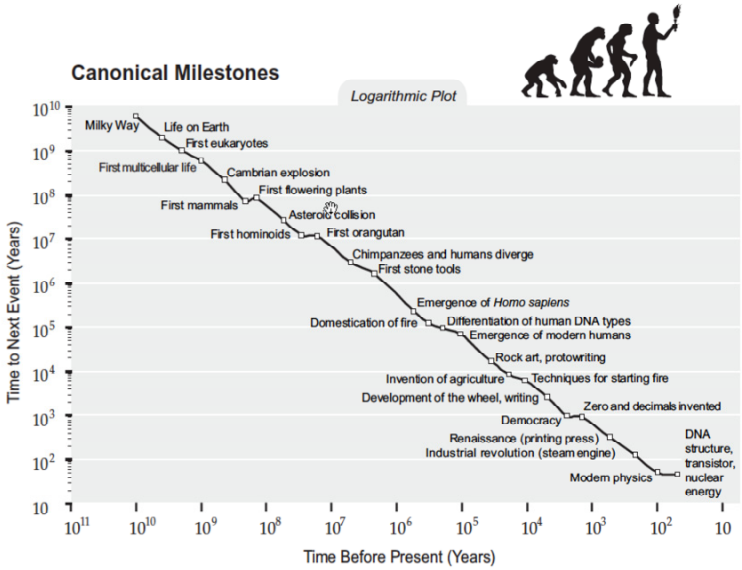
Countdown to Singularity: Biological evolution and human technology both show continual acceleration, indicated by the shorter time to the next event (two billion years from the origin of life to cells; fourteen years from the PC to the World Wide Web).

The Exponential Pace of "Progress" (cont'd)



Fifteen views of evolution: Major paradigm shifts in the history of the world, as seen by fifteen different lists of key events. There is a clear trend of smooth acceleration through biological and then technological evolution.

The Exponential Pace of "Progress" (cont'd)



Canonical milestones based on clusters of events from thirteen lists.

Epoch 5 — Merger of human technology & human intelligence

1. Cyborgs
2. Robotics
3. Cognitive prostheses

Epoch 6 — The Universe Wakes Up

Singularity · 3 — Possibility · Conceptual

1. Unless one is a dualist, it is hard to construct a compelling argument against the **possibility** of a singularity.
2. There is no physical law, after all, that precluding particles from being organized in ways that perform even more advanced computations than the arrangements of particles in human brains
3. If evolution did it, why can't we?

Singularity · 3 — Possibility · Chalmers

1. Definitions
   - **AI** ≡ human level machine intelligence (pass Turing Test)
   - **AI+** ≡ substantially greater intelligence than human
   - **AI++** ≡ much greater intelligence yet
2. Claim
   - $\mathbf{T_1}$ ≡ There will be AI (before long, absent defeaters)
   - $\mathbf{T_2}$ ≡ If there is AI, there will be AI+ (soon after, absent defeaters)
   - $\mathbf{T_3}$ ≡ If there is AI+, there will be AI++ (soon after, absent defeaters)
   - **C** ≡ There will be AI++ (before too long, absent defeaters)
3. Argument
   a) For $\mathbf{T_1}$: — Continuing success of AI
      — Brain emulation (replication?)
      — Evolution (simulated? real?)
   b) For $\mathbf{T_2}$ — If there is AI, AI will be produced by an extensible method
   c) For $\mathbf{T_3}$ — If there is AI, AI will be produced by an extensible method, *which method can itself be extended* (recursively...)

Singularity · 4 — Merits & Demerits

1. **Merits (Pro)**
   a) Eradication of disease?
   b) Eradication of poverty (through untold advances in efficiency and production)
   c) Eradication of war? (some of the triumphalists think so)
   d) Everything that civilization offers is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools that AI may provide" (from an "Open Letter on Research Priorities For Robust And Beneficial Artificial Intelligence, Future of Life Institute)*
   e) ... etc. *(more good examples are easy to imagine)*
2. **Demerits (Con)**
   a) Autonomous-weapon systems that choose and eliminate targets
      — *NB: The UN and Human Rights Watch advocate a treaty banning such weapons*
   b) Transformation of our economy for elite (oligarchical) wealth and terrible dislocation.
   c) Computers that outsmart financial markets, out-invent human researchers, out-manipulate human leaders, and develop weapons we can't understand
   d) Displace, replace, eradicate humans
   a) ... etc. *(more terrible examples are easy to imagine)*

* http://futureoflife.org/ai-open-letter/

Singularity · 5 — Reactions · To Kurzweil's 2005 Book

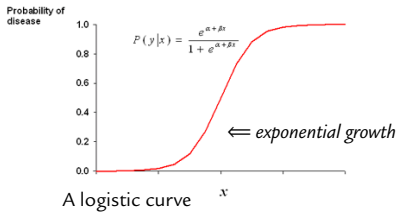
1. A "breathless romp across the outer reaches of technological possibility" while warning that the "exhilarating speculation is great fun to read, but needs to be taken with a huge dose of salt"

   — Paul Davies (in Nature)

2. An echo of "apocalyptic myths in which history is about to be interrupted by a world-transforming event"

   — John Gray

3. Many people (even if it is totally unlikely that they would put it in this language!) took the assumption of continuing exponential growth to be *daft* — thinking that it what is coming is far more likely to be a *logistic*.



A logistic curve

Singularity · 5 — Reactions · More recent

1. Over the intervening 11 years, the possibility has been taken more and more seriously, and responses are growing stronger
2. It is increasingly recognized both:
   a) That something like a singularity is a very real possibility
   b) That if/when it occurs, it may well be the biggest event in human history
   c) That we are not talking about something in the far distant future—but something that may occur within (opinions vary) something like 20–80 years
3. I.e., there is a very real chance that at least the beginnings of this transformation will occur **during your lifetimes!**
4. Some notable examples ...

Bill Joy (Inventor of Java)

1. "The fate of the human race would be at the mercy of the machines"
2. "The 21st-century technologies—genetics, nanotechnology, and robotics (GNR)—are so powerful that they can spawn whole new classes of accidents and abuses. ... [T]hese accidents and abuses are widely within the reach of individuals or small groups. .... Knowledge alone will enable the use of them. ... [W]e have the possibility not just of weapons of mass destruction but of knowledge-enabled mass destruction (KMD).
3. "I think it is no exaggeration to say we are on the cusp of the further perfection of extreme evil, an evil whose possibility spreads *well beyond that which weapons of mass destruction bequeathed to the nation-states*"
4. "[W]ith the prospect of human-level computing power in about 30 years, a new idea suggests itself: that I may be working to create tools which will enable the construction of the technology that may replace our species. How do I feel about this? *Very uncomfortable.*"
5. "We are being propelled into this new century with no plan, no control, no brakes"

Elon Musk (Tesla, SpaceX)

"I think we should be very careful about artificial intelligence. If I were to guess like what our biggest existential threat is, it's probably that. So we need to be very careful with the artificial intelligence. Increasingly scientists think there should be some regulatory oversight maybe at the national and international level, just to make sure that we don't do something very foolish. With artificial intelligence we are summoning the demon. *In all those stories where there's the guy with the pentagram and the holy water, it's like yeah he's sure he can control the demon. Didn't work out.*"

Bill Gates

"I am in the camp that is concerned about super intelligence. First the machines will do a lot of jobs for us and not be super intelligent. That should be positive if we manage it well. A few decades after that, though, the intelligence is strong enough to be a concern. I agree with Elon Musk and some others on this and don't understand why some people are not concerned."

Steven Hawking

1. "I think the development of full artificial intelligence could spell the end of the human race"
2. "Once humans develop artificial intelligence, it will take off on its own and redesign itself at an ever-increasing rate ... Humans, who are limited by slow biological evolution, couldn't compete and would be superseded."

Steven Wozniak

1. "Like people including Stephen Hawking and Elon Musk have predicted, I agree that the future is scary and very bad for people," Wozniak said. "If we build these devices to take care of everything for us, eventually they'll think faster than us and they'll get rid of the slow humans to run companies more efficiently."
2. "Computers are going to take over from humans, no question."

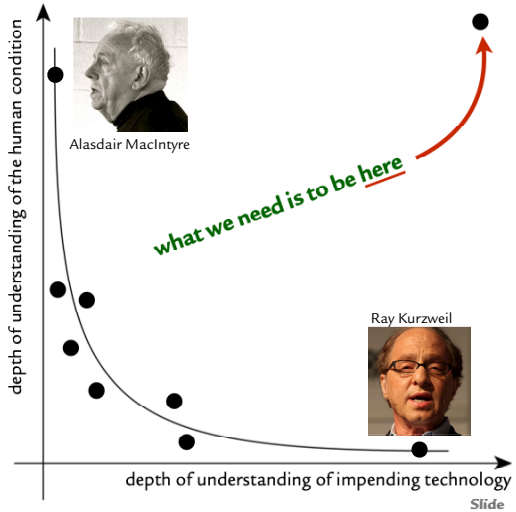
Someone else ...

We "suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and machines become more and more intelligent, people will let machines make more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machines off, because they will be so dependent on them that turning them off would amount to suicide."

Ted Kaczynski (the Unabomber)

Moral: Human reaction to the singularity (or anything approaching it) may be violent...

Singularity · 6 — Prospects ...



Singularity · 6 — Prospects ... (cont'd)

1. Flourishing
2. Extinction
3. Isolation
4. Inferiority
5. Integration

You—the students of this course—are the people who will make the decisions about, and encounter, the singularity. What happens, how it is understood, how the world reacts, is up to you ...

Thanks for coming on the trip!