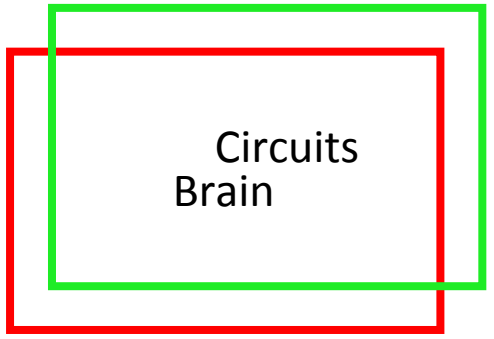
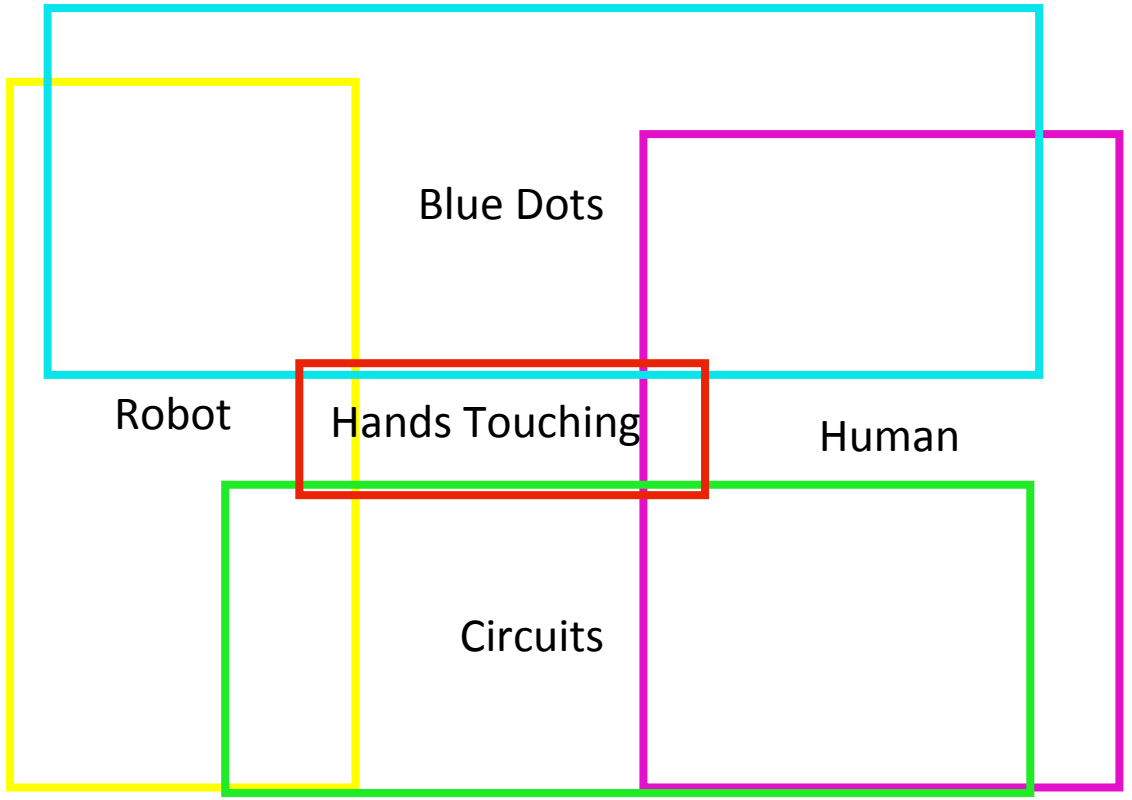
Challenges for an Ontology of Artificial Intelligence

Scott H. Hawley, Ph.D.

Associate Professor of Physics

Belmont University, Nashville TN USA





"AI Hype" Image

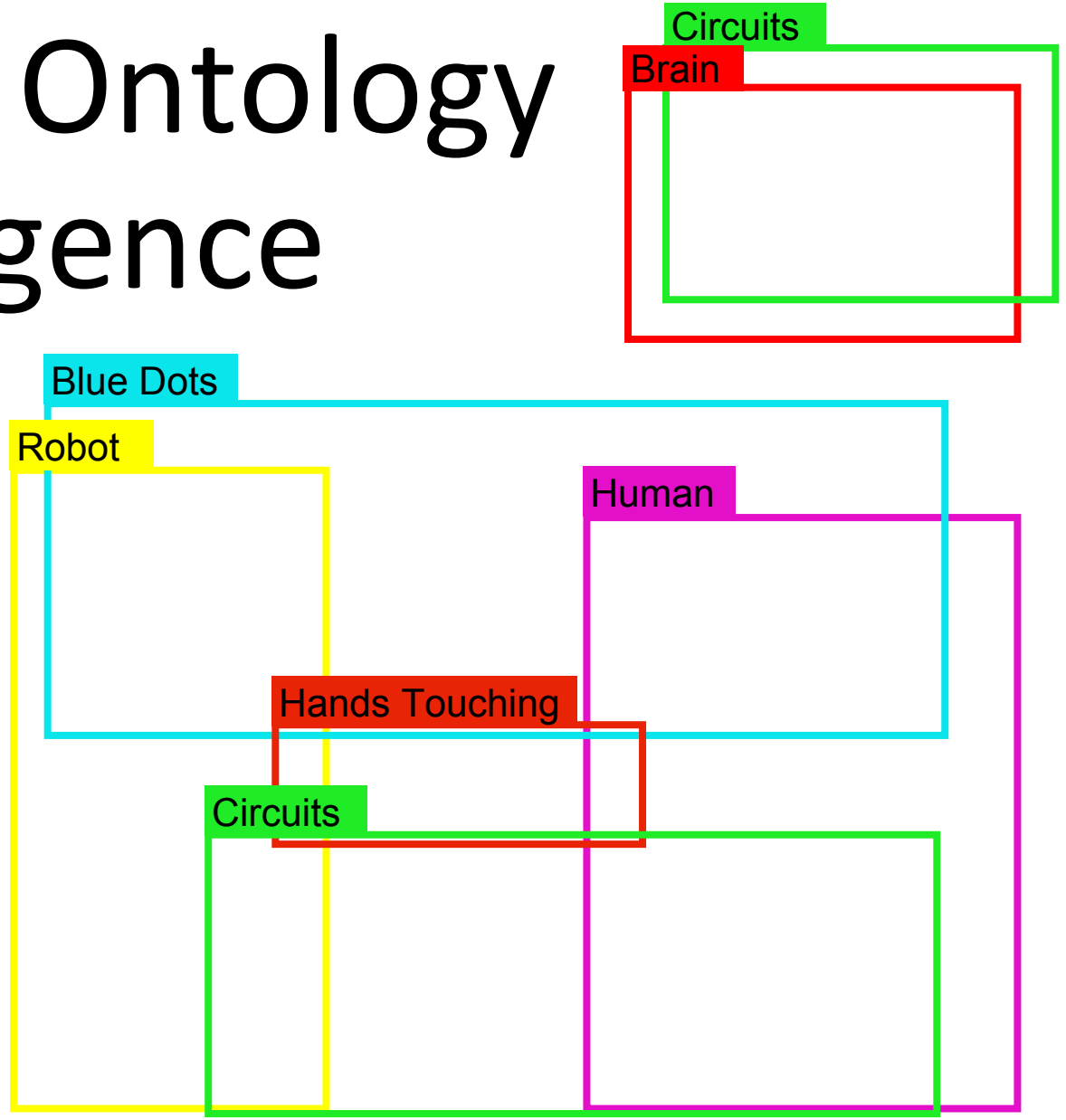
Computer Science Dept, University of Bath, August 2 2018

Challenges for an Ontology of Artificial Intelligence

Scott H. Hawley, Ph.D.

Associate Professor of Physics

Belmont University, Nashville TN USA

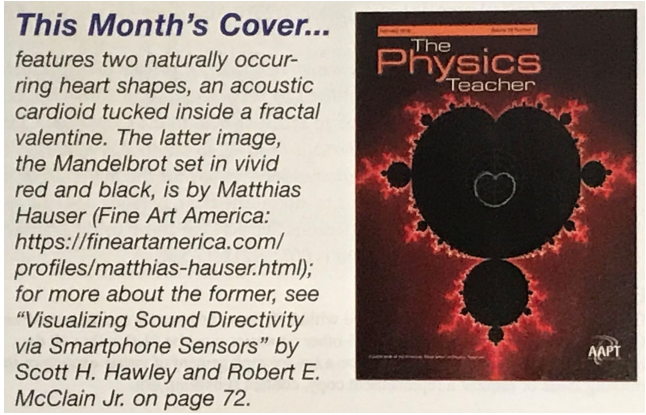


Computer Science Dept, University of Bath, August 2 2018

Speaker @drscotthawley

- Ph.D. in Physics (Numerical Relativity) from Univ. of Texas at Austin in 2000.
- 6 years as postdoc: High Performance Computing sim's of binary black holes
- Joined Belmont faculty to *teach* audio engineers and play music, switched to acoustics & audio, wrote visualization apps to help teach
- Began researching/developing machine learning for signal processing in 2015
- Started "ASPIRE: A Research Co-op" in Nashville in 2017. Community network of engineers, scientists & hobbyists

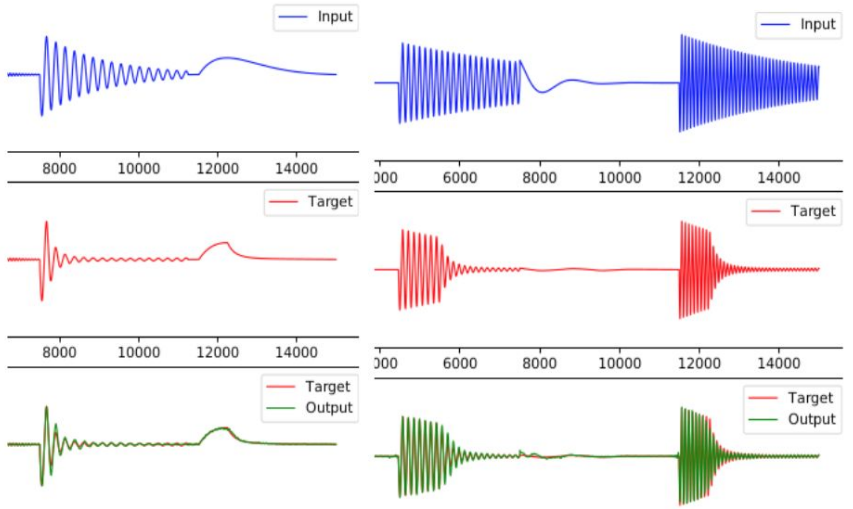
February 2018:



July 2018:

In progress, with Ben Colburn & Stylianos Mimilakis (Fraunhofer IDMT):

Compressor:



Speaker / Context 2

- Why I'm in the UK: SCIO B2C2 / Project. Hence strong "AI Ethics" interest since last summer
- Here at Bath to engage in dialog with & *learn from YOU!*
- Prepping for paper for PSCF issue on AI, CFP expressed interest in ontology:

  "Once we have established the ontological question of who we are and what machines are, we can start asking the questions about the best way to move forward, including questions about the appropriate use of AI." (Schuurman, 2018)
- Observations: My own tendency to anthropomorphize as I conduct (some) ML research.
  - Why some algorithms and not others? (I've written many iterative solvers…)
  - Not obviously due to a lack of understanding/transparency: e.g., I (re)wrote it in Excel, but *still*!

Ontology: Definition, Motivation

"Ontology" in the philosophers' sense of being & essence: **What *<u>is</u>* AI?**

- **Not** the Comp. Sci. sense, e.g. relational graphs of representations, etc.

Motivation:

- Traditionally, philosophers would say "things *act* in accordance to what they *are*"
- George Grant: " ...it has been truthfully said: *technology* is the *ontology* of the age."
- Machine learning (ML) & AI are set to become dominant technology:
  - Andrew Ng: "AI is the new electricity"

Motivation, pt. 2

- Many are asking “What *is* AI”?
- We might want to develop an ontology of AI, for addressing questions such as:
  - What 'are' these systems?
  - How are they to be regarded?
  - How does an algorithm come to be regarded as an agent?
  - (Lots more questions in Schurmann’s CFP)
- But “what is AI” seems to assume…
  - AI is a distinct object or concept that can be well-demarcated
  - AI is independent from ourselves (humans), both in definition and usage

Is Ontology of AI Necessary? Or Possible?

- *Do we even need to worry about* what AI "is", or just what it "does"?
- "Instrumentalist" perspective
  - Actor-Network Theory (Bruno Latour): All operating "actants" in a "network" are characterized on the basis of how they *affect* others, not by what they *are*
  - In Psychology: Behaviorism
  - Note "strong instrumentalism" *is ontology:* "it *is* its interactions"
- "Process" philosophy (Whitehead) is relevant, yet still an ontology
- Ontology of AI may be inseparable from larger context of Human-Computer Interaction (HCI)
- And yet, to facilitate *trust* and *safety*, the principle of *transparency in AI design* (WTB) implies that "what's inside the box" (i.e., ontology) matters.
- Not claiming that an "AI Ontology" is tractable, but if one wants to pursue it, there are (at least) a few challenges to be handled with care…

WTB= Wortham, Theodorou & Bryson, 2016-2017

Outline of 3 Challenges

Entail one or more of: actual ambiguities, opportunities for miscommunication, human tendencies that obscure/conflate, and philosophical presuppositions.

1. Various definitions of AI
2. The ‘New Normal’
3. Anthropomorphism

This list is not exhaustive

Challenge 1: Various Definitions of AI

When someone says “AI”, they mean…?

- Turing: intelligent behavior is the ability to achieve human-level performance in all cognitive tasks, sufficient to fool an interrogator. (“tasks”+”investigator” = suggests Instrumentalist view)
- McCarthy (Dartmouth conference proposal): the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it

Russell & Norvig text, Figure 1.1:

"The exciting new effort to make computers think . . . *machines with minds,* in the full and literal sense" (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	"The study of mental faculties through the use of computational models" (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act" (Winston, 1992)
"The art of creating machines that perform functions that require intelligence when performed by people" (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better" (Rich and Knight, 1991)	"A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes" (Schalkoff, 1990) "The branch of computer science that is concerned with the automation of intelligent behavior" (Luger and Stubblefield, 1993)

Figure 1.1 Some definitions of AI. They are organized into four categories:

Systems that think like humans.	Systems that think rationally.
Systems that act like humans.	Systems that act rationally.

AI Definitions: Specialized Nomenclature

Given the variety of approaches to AI, subfields and sub-topics with distinct names have arisen. These can offer clarity (when listeners are familiar with them)

- Classic AI
- Machine Learning
- Deep Learning
- Weak AI, Strong AI
- Artificial General Intelligence
- Also, DARPA's "Waves of AI" (John Launchbury)
- ….to name just a few

Gratuitous Venn Diagram:

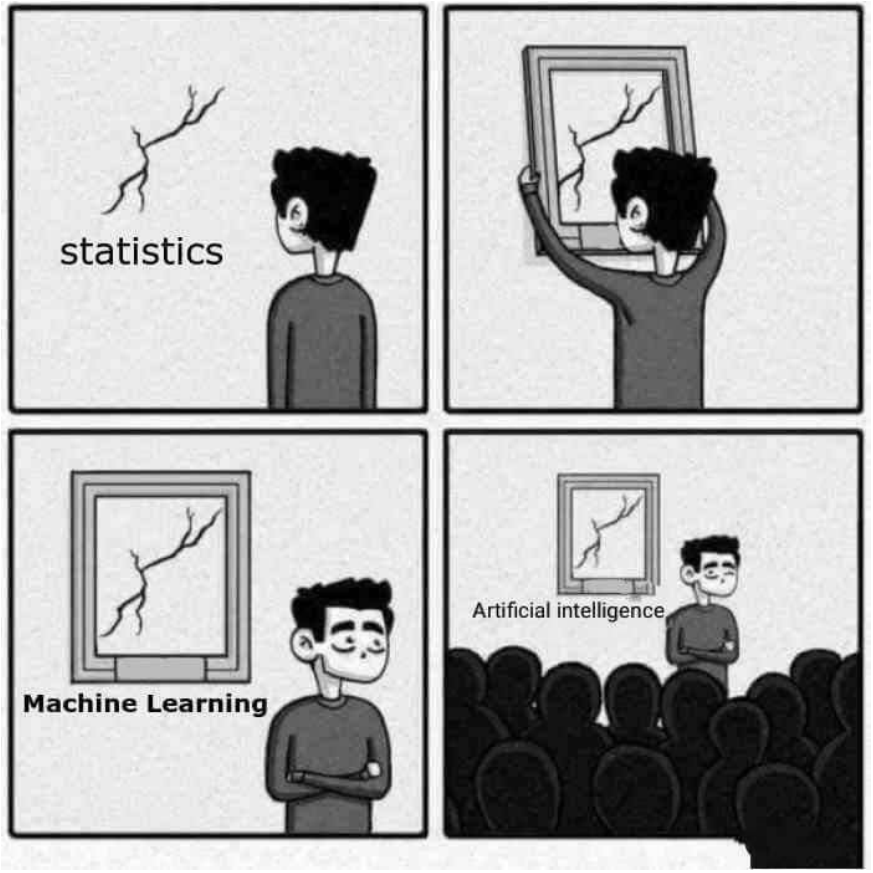
$$DL \subset ML \subset AI$$

“Classic AI”

- Characterized by human-programmed expertise & patterns
- Exemplified in Expert Systems, hard-coded decision trees (e.g. TurboTax)
    - (We’ll come to this later: I contend that not many people nowadays would regard a series of preprogrammed if-then statements as “AI” anymore)
- Systems that were programmed to play games:
    - Video game adversaries
    - Chess: DeepBlue, Stockfish,...
- ELIZA (Weisenbaum): parroted users’ inputs, passed ‘weak’ Turing test sometimes
- Note that “Classic AI” is an ex-post-facto label

Machine Learning (ML)

- AKA “Statistical Learning”
- “the study of algorithms that allow computer programs to automatically improve through experience.” -- Tom Mitchell’s text
- Iterative optimization algorithms
  - ***Long-standing numerical approaches in many fields are being rebranded as ML**
  - thus even ML definition is a bit of a moving target
- Recent characterizations (2018):
  - “Statistics at scale” – Guy Royse
  - ***“Correlation machines” – Will Geary**
  - “[Mere] Curve fitting” – Judea Pearl



--SandSerif

Within ML (i.e. “statistics” / “curve fitting”)…

- Various methods & approaches.
  - Random Forests, Hidden Markov Models, Non-negative Matrix Factorization, Independent Component Analysis, Naïve Bayes, Gaussian Processes,…,…, *and*:
- Neural Networks (NN): “linear algebra” with nonlinear operations in between
- Deep Learning
  - Neural network with hierarchical ‘layers’
  - Name was coined by Hinton to avoid stigma of “neural networks”
- Amazingly successful, domain-agnostic methods

(Aside: Plenty of ML apps don’t necessarily fit the “humanlike” part of AI)

Weak vs Strong, AGI

- “Weak AI”: task-specific human-level competencies, e.g. many recent ML successes.
  - Status: Currently exists (But do any practitioners actually say “weak AI”?)
- “Strong AI” or Artificial General Intelligence (AGI): (mimicry of) human-like performance across all cognitive domains.
  - Status: Vast body of fiction work (see “anthropomorphism”); still waiting on any code.
- My bias: Interested in ML, representations & function spaces, systems & safety, ethics, security, for “the next 20 to 30 years.” Not really AGI & future millennia.
- Andrew Ng: “AI+ethics is important, but has been partly ***hijacked*** by the AGI hype. Let's cut out the AGI nonsense and ***spend more time on the urgent problems***: Job loss/stagnant wages, undermining democracy, discrimination/bias, wealth inequality.” (Twitter, 11 June 2018, emphasis mine)

“Folklore” Definition of AI

“AI is computers doing anything we ***used to think*** only humans could do.”

-- Help: find Attribution? (aside for NPR 2018)

- This definition is remarkably effective at modeling common perception and usage in a societal context…

- The “used to” part leads us to Challenge 2 for an ontology of AI…

Challenge 2: The New Normal

Douglas Adams: “I've come up with a set of rules that describe our reactions to technologies:

1. Anything that is in the world when you’re born is normal and ordinary and is just a natural part of the way the world works.
2. Anything that's invented between when you’re fifteen and thirty-five is new and exciting and revolutionary and you can probably get a career in it.
3. Anything invented after you're thirty-five is against the natural order of things.”..."and the beginning of the end of civilisation as we know it until it’s been around for about ten years when it gradually turns out to be alright really.”

Terminology: The term “reification” is sometimes used to describe this assimilation and normalization of technology

The New Normal, re. AI

- Now that speech recognition is essentially “solved,” do ”people in general” (still) regard speech-to-text (itself) as “AI”?
- Now that systems are able to learn from “experience”, do people still regard Expert Systems as AI? Or do they say, “That’s just…”
- When one hears, *“That’s not really AI, that’s just…”* (ontological statement)
  - may indicate the speaker reserves “AI” for AGI, or
  - it may indicate a change in attitude, i.e. a re-estimation of the worthiness of the “AI” label in favor of a more specific / less *anthropomorphic* label… (two slides away)

Converse(?): Old “normal” is also now “AI”!

With “AI Hype,” we’re seeing “everything” getting labeled as AI

- As noted earlier, longstanding statistical methods are now “ML,” and therefore AI (b/c Gratuitous Venn Diagram)
- *Anything that a robot does*, now often gets regarded (by journalists & the public) as AI:
  - “Google Assistant Learned How To Fire A Gun: Should You Be Scared?” – TechTimes.com, 31 May 2018
- Recent article exploring history & dynamics of AI Hype & distortion: https://www.theguardian.com/technology/2018/jul/25/ai-artificial-intelligence-social-media-bots-wrong

Challenge 3: Anthropomorphism

The tendency to ascribe human faculties and/or intentions to entities in the world (animals, machines, objects, “forces of nature”)

- Francis Bacon observed that it often impedes our understanding of the natural world.
  - “The Idol of the Tribe: ...For it is a false assertion that the sense of man is the measure of things.” (*Novum Organum*, 1620)
- Despite its association with unenlightened eras, anthropomorphism occurs even today – perhaps even more prevalent.

  “Although commonly considered to be a relatively universal phenomenon with only limited importance in modern industrialized societies—more cute than critical—our research suggests precisely the opposite.” (Waytz et al, 2010)

Anthropomorphism & Cognition

"Unavoidable" human tendency

- "Anthropomorphization is incurable disease for human"
  - – Fumiya Iida, in "Friend in the Machine" (Beth Singler, producer)

- Operates as humans' go-to Model / Feed-forward response / Metaphor ("hammer" we try to apply to many "nails")
- Speculation: this arises because humans are hyper-social, our cognition has evolved to process our local world which is predominantly social
- May be more likely to arise for persons/situations for which detailed operational knowledge is not available (cite Wortham again?), is there unexpected/emergent behavior that seems agent-like (AlphaGo new moves)
- Is a "cognitive bias" (Wortham), and as such impedes one's ability to regard things as they are (i.e. ontologically)

Anthropomorphism in AI Design

- The earliest formulations of the concept of AI are anthropomorphic
  - Turing test (vs. Chinese Room)
  - Dartmouth conference
- 'Useful' for social robotics
  - Allow for more intuitive use
  - Can facilitate 'care' uses, e.g. w/ autistic children, some elderly care
  - Also could be 'hijacked' to create 'inappropriate' bonding (EPSRC,DP#4)

- Aside: Is "uncanny valley" a clue that ontology matters to humans (for trust)?

Effects of Anthropomorphism

- "Moral confusion" (Wortham 2018):
  - "There are serious concerns that our anthropomorphism and misunderstanding of the nature of robots extends so far as to attribute them either moral patiency, moral agency, or both (Bryson and Kime, 2011; Bryson, 2018; Gunkel, 2017b)."
- "Overidentification" (Bryson & Kime 2011): Ascribe additional human attributes based on performance at tasks of logic & language
- Can distract conversations from real, immediate dangers & opportunities, to speculations on severely underdetermined scenarios set in the far future (e.g., Ng quote 2018)
  - i.e., makes it too easy to write (yet more) fiction about AGI & waste our time

Converse (or Dual?): Dehumanization

- The tendency to regard humans as non-human entities (animals, ***machines***,...)
- Arises when…
  - Speaking of "other" groups (enemies, races,...)
  - Modeling humans for manipulation, to incite emotional responses
  - Often manifests in the form of *ontological* claims (e.g., Dennett)
- Philosophical bias: Materialism: human mind is just a 'wet' computer
  - Thus human beings *<u>are</u>* things, and not persons. ("Just collections of atoms")
- Flip-side of anthropomorphism
  - "In humanising [robots], we...further dehumanise real people" (Bryson 2009)
- Link to automation-unemployment: "The extent to which we view humans mechanistically is the extent to which will automate people out of jobs"

Mitigating Anthropomorphism

- Relevance for ontology: If anthropomorphism is a challenge, might mitigating it lessen the challenge / provide greater **clarity…**?
- Transparency (e.g., WTB 2016 & 2017, Wortham 2018)
  - “the extent to which the internal state and decision-making processes of a robot are accessible to the user” (Wortham & Theodorou, 2017)
  - “Components” should be exposable (way for users to "lift the curtain" - EPSRC PoR)
  - bot should manifest itself as (and/or declare itself to be) a bot?
  - Wortham(&T&B) found human subjects’ accuracy, i.e. **clarity**, in identifying robot’s mental model **improved** when bot’s decision processes were shown
- Note: NNs are generally *not* transparent (& not GDPR compliant)
  - Exposable vs. Explainable: Does a complicated image of layer activations = Transparency?

Moving Forward, 1: Human O. <--> AI O.

Observation: Discussions on the nature of AI are typically accompanied by discussions on the nature of humanity -- “i.e. what does it mean to be human?”

- (...and how “AI” is like or unlike “human”)

- This can be regarded as a discussion about ontology
  - Human ontology <--> AI ontology
- Or not! Attributes & properties: are they *necessarily* ontological, or can they be regarded purely functionally (instrumentally)?
- Dehumanizing assumptions preclude this productive avenue

(Since my grant is from Templeton, this area might be the key point at which religious perspective(s) could be relevant.)

Moving Forward, 2: More Answers to “What is AI”

In Machine Learning context, AI is...

- Encoded Bias / “Stereotyping at Scale”
- Classification as Power (Kate Crawford)
- A means -- to what end? Paraphrasing line from Sherry Turkle: “What does a ML system *want*?”
  - “Just a tool” language is evidence of unreflective denial
  - “Data-hungry algorithms make for data-hungry companies” (SH)
  - “A heat-seeking missile has a goal” (FHI)

In AGI context:

- “The greatest existential threat humanity has ever faced”?
- Bostrom: “What does a superintelligence want?”

In Closing...

- We have not established an ontology for AI, just pointed out challenges involved in doing so.
- The language we use is important because it belies our ontology
  - “That’s not AI, …”
  - Who is the “we”? Engineers, public,...?
  - Eventually, our language influences our ontological commitments
  - Using alternate/specific terminology helps, and yet “AI” sells.
- Key Question: Does an ontology of AI “get you anything” that an instrumentalist perspective doesn’t?
  - Indirect Answer: “Ontology creep” & reification
  - Bath group’s own efforts involve assertions that seem to be based on at least minimal ontology of AI != human (e.g., bot should be manifest-ably non-human, bots aren’t responsible/shouldn’t have rights). Can instrumentalism support these?

Further Discussion: What is the goal of AI development?

Depends on *who* you ask:

- "We want to make 'Her' like in the movie 'Her'" (ML Ph.D. on Twitter)
- Use it to solve various problems (w/o any 'consciousness')
- Model to better understand the human mind/brain

Related: *What is the "reward function" of society's current ML/AI research development & deployment enterprise?*

- If "making a better world," by whose standards?
- If "profit," that leads to a predictable set of outcomes. (Replace "profit" with "paperclips",...)
- USA Dept. of Defense (July 26 2018): "We want to be the threat"

Aside: Jacques Ellul: successful tech. development tends to become self-serving

Acknowledgement & Thanks

- Michael Burdett & Alister McGrath (both @Oxford), Tommy Kessler (Belmont), Beth Singler (Cambridge), and this group at U. of Bath!
- Belmont University
- Sponsored by a grant given by Bridging the Two Cultures of Science and the Humanities II, a project run by Scholarship and Christianity in Oxford (SCIO), the UK subsidiary of the Council for Christian Colleges and Universities, with funding by Templeton Religion Trust and The Blankemeyer Foundation.