

IACAP 2019 Notes

Mexico City, Mexico

David Abel*
david_abel@brown.edu

June 2019

Contents

1	Conference Highlights	3
2	Wednesday June 5th	4
2.1	Fiona McEvoy on AI-Driven Behavior Change	4
2.2	Karen Gonzalez-Fernandez on Logic and Heuristics in Cognitive Psychology	5
2.3	Keynote: Alexandru Baltag on The Topology of Surprise	6
2.4	David Fernandez-Duque on Stratified Evidence Logic	10
2.5	Armando Castaneda on Distributed Computing with Bounded Rational Agents	12
2.6	Fernando Velazquez-Quesada on Reliability-Based Preference Dynamics	13
3	Thursday June 6th	15
3.1	Björn Lundgren on Self-driving Cars: An Ethical Overview	15
3.2	Steve T. McKinlay on Machiavellian Machines: Glitching AI	17
3.3	Ronald Arkin on Misdirection in Robot Teams	18
3.4	Covey Award Talk: John Weckert on Our Technology Fetish	19
3.4.1	Background	20
3.4.2	Do we have a tech fetish?	20
3.4.3	Quotes on What Computers Should Do	21
3.4.4	Returning to the Fetish	23
3.4.5	Erlant Etxeberria on Mathematical Explanations and Counterpossibles	24
3.5	Anders Herlitz on Predictive Fairness	26
4	Friday June 7th	29
4.1	Alan Wagner on Kantian One Day, Consequentialist the Next	29
4.2	Angelica Fleury on Creating Moral Ideals in Robots from Buddhism	30
4.3	Simon Award Talk: Juan M Durán on Philosophically Analyzing Simulations	32
4.3.1	View 1: Problem Solving Techniques (PST)	32
4.3.2	View 2: Description of Patterns of Behavior (DPB)	33

*<http://david-abel.github.io>

4.3.3	Philosophical Implications	33
4.3.4	Conclusions	34
4.4	Alan Wagner on The Aristotelian Robot	35
4.5	Himavath Jois on Should Robots be Allowed to Punish Us?	36
4.6	Presidential Address: Don Berkich on Underdetermination after Computation	38

This document contains notes I took during the events I managed to make it to at (my first) conference of the International Association for Computing and Philosophy (IACAP), in Mexico City, Mexico. Please feel free to distribute it and shoot me an email at david_abel@brown.edu if you find any typos or other items that need correcting.

1 Conference Highlights

IACAP was a really fascinating experience—I came away with a mix of thoughts on the interplay between philosophy and computing, and the role computing is playing in our society at the moment (questions, I find, distinct from those typically raised at ML conferences).

One thing I loved about the conference was the scope of material: almost any topic/question was on the table to be explored. It’s one thing that fundamentally separates CS from philosophy: philosophy is (often) concerned with asking questions in the right way and clarifying our inquiry, no matter the topic. ML/AI, at least at the moment, is usually focused on specific technical contributions (often a new algorithm, new theorems, experiments, and so on). Of course, these different artifacts serve different purposes, and both matter a great deal. To this end, the conference spanned a great deal of material, from non-classical logics, to self-driving cars, to meta-ethics, to the role computation plays (or should play) in the philosophy of science.

A few takeaways:

1. Philosophy and computation (and, to me, specifically ML/AI) still have loads to share with one another.
(! But, doing so can be very challenging: different tools, methods, and objectives!)
2. I really enjoyed John Weckert’s talk “Our Technology Fetish”, discussing the world’s current irrational obsession with development of technology (even in cases where we don’t have an obvious use for such systems). In many ways it was a call for rethinking some of the values of what constitutes a flourishing society.
3. Fiona McEvoy’s talk also hit home: she discussed how automated systems are driving behavioral change through “nudges” in online (and other settings)—it had flavors of Zeynep Tufekci’s 2019 ICLR keynote. The message in both cases is for us to be wary of the kinds of psychological impact that subliminal interventions online can have on our communities, such as *fundamental* changes to our worldview and values.
4. I *loved* how international and friendly the conference was: it’s a tight knit community of folks that have been attending IACAP for awhile, so many have known each other for quite some time. It also drew in people from all over the world, so the perspectives shared about the role of technology were extremely diverse.

2 Wednesday June 5th

It begins! Today is primarily a symposium on logic and computing, so the latter half is more focused on those topics (though I had to step out for the second half of the day).

2.1 Fiona McEvoy on AI-Driven Behavior Change

Goal: recommendations for future research on *decision guidance* algorithms (not decision-making algorithms).

! Example: Google results based on clicks/habits, recommendations systems.

“Nudge”: any aspect of a system/intervention that alters people’s behavior in a predictable way without forbidding any options or significantly changing economic incentives. Think: bench with bars on it vs. a sign.

Two Decision-Making Systems: System 1 (automatic/intuition), and System 2 (rational thinking). “Nudge” is making use of System 1.

Definition 1 (Hypernudging): *A feedback loop online that contributes heavily to how individuals make decision.*

! Data collection about decisions and actions.

Point: Hypernudging has been accused of manipulation.

Q: Which features keep nudges consistent with autonomous human decision-making, and which force them to be manipulative?

Hypothesis: the more imersive an environment is, the more they can manipulate individuals.

Definition 2 (Autonomy): *Two notions: 1) the opportunity for autonomy; Mill's freedom from manipulation, and 2) capacity for autonomy; I should be able to pursue my goals.*

One group: The environment is full of non-rational influences. So, even our most rational decisions fall out of non-rational effects. Therefore, it must be permissible to shape the choice environment.

! Distinction: shaping environment vs distorting! “Shape” is [Dave: I missed it, something like “A intends to Y, based on Z”](#), vs. distort: “A intends to do X”.

Features of ethically-permissible nudges:

A range of options from which to choose

Can make a decision that differs from the preferences of the individual.

Environment is non-deceptive/distorting/subliminal.

Three case studies:

1. Online (based on Karen Yeung's work on hypernudges [6]).
/ Yeung says most online systems are manipulative in the sense that people are coerced into making non-rational decisions.
2. Voice Controlled IoT: Whoever controls these has immense power: company that makes these software/hardware systems gets to choose a lot of things for you (which music service you use, shopping platform, and so on).
/ Relationship between IoT/voice systems and people is becoming richer and more complex. Relationships/sensitivity develops, which masks the fact that these items are commercial entities.
3. Virtual/Augmented Reality: New frontier – nudges/distortion possible by visual/3d construction of avatars, lots of opportunity to manipulate users through the environment.

/ Example: put elementary school kids into a VR environment, and they *remembered* the experience (1-2 weeks later) as a real experience (not as VR?).

By the nature of all three environments, there is heavy psychological potency to manipulate users.

Wired: “AR will spark the next big tech platform.” / Claim: we will use AR as the next frontier.

.....

2.2 Karen Gonzalez-Fernandez on Logic and Heuristics in Cognitive Psychology

Logic and heuristics have a long history: heuristics tends to be linked to the process of discovery, particularly under uncertainty and economical considerations. Conversely, logic is about deduction and certainty.

Q: Should logic and heuristics both be considered as part of the norms of reasoning in people?

Goal: Show how logic and heuristics are understood as criteria of rationality.

Q: What is the role of logic/heuristics in rationality in cognitive psychology?

Proposal 1: “To be rational is to reason in accordance with principles of reasoning that are based on rules of logic, probability and so forth...such rules are normative principles of reasoning” [4].

Proposal 2: heuristics are “strategies that allow us to make plausible inferences economizing cognitive resources”.

/ Kahneman and Tversky [3]: two systems of reasoning.

1. System 1: intuitive system

2. System 2: rational system.

/ Heuristics appear in judgments of system 1 because they are more accessible.

/ Heuristics produce **bias** and must be studied mainly to overcome/strategise w.r.t. systematic errors we make.

Gigerenzer: proposal for ecological rationality [5]. In “small worlds” only classical logics can be applied. In “large worlds”, logic can’t be applied, so we need heuristics.

/ Can develop either descriptive models or normative models of heuristics.

Proposal: Logic vs. heuristics as a criterion of rationality. Consider the following properties:

Logic	Heuristics
Logical Omniscience	Lose valid inferences
Infallibility	Fallibility
Consistency	Inconsistency
Context-free rules	Rules influenced by context
No constraints	Limitations of time/memory

/ Takeaway: logic and heuristics are in conflict with one another.

/ But, from recent work in AI, perhaps we can find a more cooperative model that incorporates both strategies. Some proposals define an algorithm as logic w/ control, that unifies aspects of both approaches.

Conclusion: Logic and heuristics can cooperate. Take Gillies proposal that Inference = Logic + control, we can:

1. Introduce non-classical logics
2. Find a better understanding of processes we make when we reason in ordinary life and inside scientific methodology.
3. Open Problem (for psychologists): Why should we maintain classical logics as a paradigm for rationality? Should we include non-classical logics in our theories of rationality?
4. Open Problem (for philosophers): What are we doing when we do logic? How should we understand the relationship between logic and reasoning?

.....

2.3 Keynote: Alexandru Baltag on The Topology of Surprise

Q: Knowing the world: Can I ever know the state of the world?

A: Well, it depends! On three things:

1. The actual world: some worlds are knowable, some not.
2. What can I observe? Topology of observable evidence.
3. Background information: are we given any prior knowledge?

! Knowability of the world sounds metaphysical and unapplied. But, in a state-space formalism, the state of the world is an abstraction: the most refined description of the world that is relevant for the given purposes.

! This description *consists of the answer to relevant questions*.

Central Question: is it possible to learn the answer to some (relevant) question, given enough observable evidence.

Consider the following paradox:

Definition 3 (Surprise Exam Paradox): *A student know the date of an exam is next week. Doesn't know which day.*

Teacher announces that the exam's date will be a surprise: even in the evening before the exam, the student will not be sure that the exam is tomorrow. (also all participants can't lie).

But, applying backwards induction: if the announcement is true, then the test cannot take place any day of the week! So, the test cannot be a surprise and the student dismisses the announcement. But then, the test will be a surprise! Paradox.

Note that there are multiple interpretations here: the student might not be able to eliminate Thursday (and the rest of the days, only Friday). That is:

1. Non-self-referential interpretation: teacher meant “you wouldn’t know in advance the day of the exam, without any help from me (if you are not using even the information that I am announcing now).
! Thus, the elimination process stops here.
2. Self-referential interpretation (most common): teacher meant “you will not know in advance the exam day, period, even after hearing this announcement.
! But, this version leads to a paradox (a non-lying teacher can’t answer the question “after hearing this announcement, will the exam still be a surprise?” five times (“no”, specifically), and avoid contradiction.

Definition 4 (Topology): *A topology on a set X is given by a family $\mathcal{T} \subseteq P(X)$, or “states”: possible descriptions of the actual world.*

So: the set of possible worlds is defined by the space X .

! Open sets represent the agent's evidence about the world: at world $x \in X$, every open set $U \in \mathcal{T}$ with $x \in U$ is a piece of evidence.

Two interpretations: 1) evidence in hand (topology of actual evidence), and 2) evidence out there (topology of potential evidence).

Q: Why a topology?

A: Start with properties of the world that are directly observable, they form a topological basis.

Typical assumptions:

CS/Economists tend to assume that the topology given is a *partition* of the state space X .
! Assumes absolute certainty.

Others tend to latch onto *potential* evidence, which leads to making no restrictions on the topologies.

Setting: multiple agents, multiple perspectives. This requires multiple topologies.

Some definitions:

1. Neighborhood of a point $x \in X$ is any open set U where $x \in U$.
2. Interior point of a set $A \subseteq X$ is a point x such that there is a neighborhood of x where $U \subseteq A$.
3. A limit point of a set A is a point $x \in X$ s.t. every neighborhood of x contains a point $y \in A$ with $y \neq x$.
4. Interior of A is the set of all its interior points:

$$Int(A) = \{x \in X \mid \exists U \in \mathcal{T} (x \in U \wedge U \subseteq A)\}$$

5. Closure of A is:

$$Cl(A) = \{x \in X \mid \exists U \in \mathcal{T} (x \in U \wedge U \cap A \neq \emptyset)\}$$

6. The (Cantor) derivative of A is the set of its limits points:

$$d(A) := \{x \in X \mid \exists U \in \mathcal{T} (x \in U \wedge (U \setminus \{x\}) \cap A \neq \emptyset)\}$$

Note: the interior satisfies the dual axioms, corresponding to a modal system (S4):

$$Int(X) = X, \quad Int(A) \subseteq A \quad (1)$$

$$Int(A \cap B) = Int(A) \cap Int(B), \quad Int(Int(A)) = Int(A). \quad (2)$$

Epistemic interpretation: captures main properties of a natural concept of knowledge or knowability (based roughly on the modal system S4).

Important Distinction: 1) you know/do not know, vs. 2) you can know/cannot know.

Example topologies:

1. Complete ignorance: the trivial topology $\tau = \{\emptyset, X\}$.
2. Omniscience (god): $\tau = P(X) = \{Y \mid Y \subseteq X\}$.
3. Knowledge based on measurements of points on a line.

So far: we have a proper definition of knowledge and knowability. Let us return to the question:

Q: Is the actual state of the world known?

A: Depends on evidence. Suppose background information is given by a subset $A \subseteq X$ of a topological space.

! Suppose: actual world x belongs to A . Then, I can know actual world x iff x is isolated in A , that is, iff $\exists U \in \tau_X$ such that $x \in U \subseteq A$.

Example: Policeman and the Speeding Car.

Policeman uses a radar with accuracy ± 2 mph to determine whether the car is speeding in a 50mph zone.

Radar shows 51 mph.

$X = (0, 1)$ is the set of possible worlds where we assume the car is known to be moving.

The property “speeding” is knowable, but it is now known in this context.

Example: Teacker marks a point x on a real line R . Announces that the point is in the set:

$$A = [0, 2] \text{ OR } [1, 2] \tag{3}$$

$$\left\{ \frac{1}{n} : n \in \mathbb{N}, n \geq 2 \right\} \text{ OR } \tag{4}$$

$$\left\{ \frac{1}{n} + \frac{1}{n^m} : n, m \in \mathbb{N}, n \geq 2 \right\} \text{ OR } \tag{5}$$

$$[1, 2]. \tag{6}$$

Q: Can the student know the position? Depends on further measurements, their accuracy, and so on.

! We can bury a surprise exam paradox in here, too.

Let's revisit a topological epistemic analysis of the surprise exam paradox: what is the potential evidence in this context?

Under the self-referential view, we can recreate the paradox. The topological analysis gave us some justification as to *why* this is a paradox: the only perfect subset of A is the empty set!

Multi-agent example: Two numbers, one on each of two agents' foreheads. The numbers are off by one.

! Ask each agent: do you know the number on your head? First time: no! Ask again: no! ask again: yes! Can use Cantor's derivative method to repeatedly eliminate possible worlds until x is known (the true state of the world).

.....

2.4 David Fernandez-Duque on Stratified Evidence Logic

Example: Alice wants to know if the movie "Roma" by Alfonso Cuaron is a good movie. She asks her chatbot, Bot. Bot decides that wikipedia and rotten tomatoes are not sufficient for answering Alice's question.

Bot does conclude that an oscar nomination and a rotten tomatoes score of 90% or higher are sufficient to answer Alice's question.

! Both can only answer her question with $N = 2$ pieces of evidence. But what if N is really large? $N > 10,000$?

Goal: Present a logical framework in which the amount of evidence needed to conclude a new belief can be explicitly counted in order to model *bounded rationality*.

Definition 5 (Evidence Models): *An evidence model consists of:*

1. *Possible worlds where atomic facts may attain different truth values.*
2. *Evidence sets: sets of worlds that remain possible after making an individual observation.*

Some evidence-based situations:

Ap : p is true everywhere

Ep : Some piece of evidence supports that p .

op : Some factual piece of evidence supports that p .

p : There is some finite collection of factual evidence supporting p .

$[\alpha]p$: there is a collection of size less than $\alpha + 1$ that supports p .

! This is the main contribution of the work: a new theory for clarifying the size of evidence and its role.

Definition 6 (Stratified Evidence Model): *A stratified evidence model is a triple, $M = (W, (E_\alpha)_{\alpha \in \mathbb{N}}, V)$, where the evidence all has to be arranged as less than a particular size/amount (as denoted by α).*

Stratified Evidence logic:

1. Syntax: $\phi, p, \psi, [\alpha]\phi$, where α is a cardinal/natural number, p is an atomic proposition.
2. New piece: what happens to stratified evidence? $(M, w) \models [\alpha]\phi$ if there is X that support ϕ subject to the α constraint.

Back to the Roma Example: Bot needed to check Rotten Tomatoes and Wikipedia to verify that Roma is a great movie. This can be represented in a strict stratified evidence model.

Example 2: Let's consider graph problems, like determining whether a graph contains a hamiltonian cycle.

Q: How might we express this problem in stratified epistemic logic?

A: Well, we pose statements like $M \models h \wedge [1]h$, if one piece of evidence can let us answer the decision problem. More generally, we'd see $M \models \neg h \wedge [\omega]h$, where ω is the first infinite cardinal.

Example 3: The halting problem. H_n is the set of machines that halt in at most n steps, W is the set of all turing machines. Then we can express this: $M \models h \wedge [\omega]h$. To determine if they don't halt, though, we of course need to write $M \models \neg p \wedge [\omega_1]p$, where ω_1 is the first uncountable ordinal (since we would have to run the machine forever).

Main result:

Theorem 2.1. *The following are equivalent:*

1. ϕ is satisfiable over a stratified evidence model
2. ϕ is satisfiable over a strict stratified evidence model
3. ϕ is satisfiable over a finite stratified evidence model.

Expressing standard evidence notions:

1. $A\phi \equiv [0]\phi$: ϕ is true everywhere
2. $E\phi \equiv \neg[0]\neg\phi$: There is evidence supporting ϕ .
3. $F\phi \equiv [1]\phi$: There is factual evidence supporting ϕ .

Concluding Remarks:

Stratified evidence logics allow us to explicitly quantify the amount/cost of evidence required to justify a given fact.

This logic retains the nice computational behavior of standard evidence logics.

Q: What is precise computational complexity of stratified evidence logic and its fragments?

Many familiar evidence-based notions naturally fall into our framework.

Evidence may also be stratified by trust rather than effort, leading to very different logics.

.....

2.5 Armando Castaneda on Distributed Computing with Bounded Rational Agents

Goal: Think about distributed systems from an agent-centric view.

Two components of a distributed system:

1. Agents: computers, robots, people, ants.
2. Communication medium: messages, signals, public/private announcements.

! We have loads of models, all of which make different assumptions; timing assumptions (synchronous/asynchronous), local failure assumptions, static/dynamic, and so on.

Q: What about *local* computations?

A: Two views:

1. Computability approach: each agent is “more” than a Turing machine
! can be useful to clarify impossibility results. But, suffers because some solvable tasks are not implementable in reality, period.
2. Algorithmic approach: TM with unbounded time/space, often a loose definition.
! People do not like locally expensive solutions (usually assume agents can solve NP-Complete problems locally).

New Proposal: Incorporate realistic models of local computations in the study of distributed systems.

! Bound resources like time, space, and bandwidth.

Compact Local Streaming (CLS) Model: Agent is a turing machine with *compact space*, where space $\propto \log k$, where k is the number of agents in the system.

Communication: synchronous failure free (SFF):

1. Communication graph with arbitrary topology.
2. Message-passing to neighbors

3. Synchronous communication
4. Nobody knows the whole graph
5. Failure free
6. Execution/communication proceeds in discrete rounds.

Target: simple and general framework for studying these notions.

Local Streaming Algorithms: each round, agents 1) send message to neighbors, 2) receive messages from neighbors, 3) perform local computations.

! Because of the space limitation imposed on agents, must treat step (2) like a streaming algorithm.

Two parts to each algorithm: 1) local streaming algorithm for reacting to messages, and 2) compact communication protocol.

Positive Results: Lots of possible results on solving various problems using this model: DFS, BFS, leader election, $\Delta + 1$ graph coloring, k -th largest element, compact routing, and so on.

Negative Results: Recall that the size of local state is bounded, size of message is bounded, and the memory of the whole system is bounded. So, there are loads of problems that can't be solved. CLS has essentially the power of a TM with poly space. So, problems that require exponential space cannot be solved in CLS.

Future Directions:

1. Continued analysis of CLS, complexity separations, epistemic logics, and so on.
2. Resource-bounded agents in other classic models.
3. Impact of CLS to classical results.

.....

2.6 Fernando Velazquez-Quesada on Reliability-Based Preference Dynamics

Main Question: How do other people influence us? More precisely, how do the preferences and beliefs of other people influence our individual preferences and beliefs?

Proposals:

Reaching a consensus from DeGroot [2]: agents have some probability distribution and a weight for other agents.

! My revised distribution is some linear combination of everybody's distributions.

Dynamics of Peer Pressure

Threshold Models

! Lots of perspectives on this phenomena. Generally, we want to understand this process.

This proposal: *reliability-based* preference change.

Q: How do other people's preferences affect ours when you have a priority ordering among them?

Setting: Finite set of agents, each with a preference relation over situations and a reliability relation over agents.

Goal: provide a function from everybody's preferences and a priority among them to a single preference relation using a *lexicographic rule*.

Approach:

1. Lexicographic rule
2. Model operation for simultaneous preference update
3. Modality describing the effect (recursion-axioms based axiom system)
4. Some variations of the operation
5. Identifying situations where the iteration leads to unanimity/stability.

Definition 7 (PR Frame): *The i -th agent will be associated with a tuple (W, f_i, r_i) where:*

1. *W is a finite set of worlds (that agents will have preferences over).*
2. *f_i : a preference relation (an ordering on worlds)*
3. *r_i : A reliability relation (ordering over other agents).*

Also need a language, semantic interpretation, and axioms for describing the worlds/models (it's roughly first order logic, with the added machinery to express preference/reliability relations).

Example 1 (Total order case): Give precedence to the most important ordering/agent. Can generate new preferences.

Example 2 (partial order case): Solutions already in literature, this formalism can roughly capture these formalisms.

! Can also handle the preorder case.

Dave: Had to run for the day!

3 Thursday June 6th

Day two! The morning session will mostly be on AI.

3.1 Björn Lundgren on Self-driving Cars: An Ethical Overview

Self-driving cars: discussion usually centers around ethical crashing and trolley problem. Two issues: 1) How should we crash (machine-ethics), 2) who is blameworthy (responsibility).

This talk: let's broaden the discussion!

Q: Why should we have self-driving cars?

A: Safety! Seems to be a necessary condition, in fact.

! But, what is traffic safety? One answer: absence of accidents. Another: absence of severe accidents!

Example: four way stops yield fewer accidents, but more severe accidents; roundabouts yield more, but less severe.

Q: How can we measure safety?

A: Karla and Paddock (2016) suggest we need a huge amount of data to *determine* whether cars are safe.

A: Alternatives? Simulations, proofs, verification, software alongside people (like Tesla).

A: Safe compared to what? Current drivers?

Fact: 93% of accidents are human-caused. Out of these: 30% speeding, 30% intoxication, and 20% distracted drivers. So, why not just solve those problems?

! Alcolocks force drivers to be sober to turn the car on, and so on. Why not just adopt these techniques.

Q: But will this actually fix the problem?

Safety beyond self driving cars:

1. infrastructure is crucial! Do we plan for SDC or human driven cars?
2. Plans require long time horizon: mixed traffic? only SDCs? Increase/decrease of road vehicles?

Other values:

How do we balance safety against other traffic values? (Like efficiency)

Is there a minimal safety level? Should we get to pick a safety level (within reason)?

If safety is increased by SDCs, does that effect our evaluation of a requirement for a minimal safety level?

Might we ban driver's license? Ban less safe cars?

Security: how do we protect against manipulation of sensors (analog hacking)? ! One option is data verification by other sources ! Might be possible with available technologies. Q: What about digital hacking?

Further issues of *privacy and autonomy*:

Information dependency raises privacy concerns.

What kind of data can we use? In car behavior of individuals?

Who can use this data? Manufacturers? Government? And so on.

How can this data be used? For safety? For commercial reasons?

Can a user opt out? What about people not in the car?

Climate and environment:

It will be so convenient to travel we will likely travel *more*.

Will we have more or less vehicles?

Will SDCs compete with public transportation?

Will SDCs be electric? Increase in demand for nickel/lithium?

Health Issues:

Walking less

Loss in organ transplant donors, since many currently come from the result of traffic accidents.

Jobs and economy:

In Europe: 4.5% of the workforce is in transportation. Huge loss of jobs.

Will we be making more or less cars? Car industry is huge at the moment.

Will SDCs increase efficiency of transport system, yielding potential for new services?

Will the economy of transportation be more of an information economy, with companies like Google dominating?

Potential for Political and Social Resistance:

Labor resistance

Economical models might change, resistance from traditional car manufacturers.

Fear of new technology (3/4 Americans fear using fully autonomous SDCs)

Argue that we have a right to drive? A natural right?

Distribution of responsibility:

Who is responsible for achieving safety and other values (like privacy)?

Who is responsible when things go wrong?

Volvo and Audi declare responsibility for accidents (but not the other values).

Law enforcement:

What constitutes reasonable search?

What is suitable balance between surveillance and privacy?

3.2 Steve T. McKinlay on Machiavellian Machines: Glitching AI

Focus: Algorithmic opacity and trust.

Main perspective in literature: do we accept the current degree of opacity, or should we expect full explainability?

This focus: more philosophical question about opacity.

! Main idea: new definition of epistemic opacity for AI.

Definition 8 (Epistemic Opacity): *A process is epistemically opaque to P if it is impossible for P , given the nature of P , to know all of the epistemically relevant elements of the process (Humphrey 2004)*

But, has some issues. So, a new extended (and stronger) definition:

Definition 9 (Emergent Opacity): *A process is emergently opaque to P if the opacity emerged as a result of some machine learning process and it is impossible given any nature of P , for P to know all the epistemically relevant elements of the process.*

Intuitive: algorithms that exhibit emergent opacity are those that arise from machine learning.

Example: Generative Adversarial Networks (GANs). Can generate images of fake people.

Problems!

Lots of clear ethical implications.

But, philosophical implications as well: deep learning suffers from epistemological holism.

/ Quine: There cannot be an (epistemic) statement of a theory that is in a vacuum.

/ (Inspired by Dummett): In the future, there's no way to determine if artifacts are computer generated or not.

Point: We're almost at an age where we can't really trust anything digital (images, videos).

It's clear: machines have a profound ability to shape our behaviour.

Glitches: often minor faults or weaknesses in a system, but are not so damaging so as to prevent individuals from using the system. In gaming community they're often embraced, studied, and exploited. Can be thought of as an opaque bug.

“Whosoever desires constant success must change conduct with the times.” Dave: (Machiavelli, presumably?)

One approach to understanding AI systems: treat it like a typical scientific study. Test hypotheses, reach conclusions. However: problematic because of the time horizon. Dave: Is it?

Alternative: an algorithmic Laplacian demon. Design algorithms that can shed light on their own internal workings.

/ (Veshala Polensky): Examined this approach (building explanatory systems) based on case studies. Identified a problem: systems designed to be persuasive end up carrying out manipulative/persuasive explanations. Polensky illustrates three cases where these kinds of manipulations take place.

.....

3.3 Ronald Arkin on Misdirection in Robot Teams

Deception: commonly used by animals (see: mimicry in butterflies, camouflage of lizards). Use of deception by primates indicates a theory of mind [1], could be a hallmark of social intelligence.

Q: How can we get humans and robots get along? Use/understanding of mimicry, camouflage, and deception might be important.

/ Happens in people! People make use of deception in sports, military, education (encourage a baby to eat vegetables by showing them a Kitkat), medical therapy.

Denett: “another price you pay for higher-order intentionality is the opportunity for deception”.

Claim: More intentional and autonomous social robots become possible given deception. Has been used in a robot referee, rock-paper-scissors, tends to increase humans' opinions of the robot (along certain axes).

The Phenomena of Deception: Bond and Robinson’s definition of deception (“a false communication that tends...”) implies five steps.

New results: A situation’s location in interdependence space can be used to determine if a robot or agent should act deception.

! Received dramatic reactions from the media, chastising the work (despite the paper containing clear statements of the nature of the work: focused on social robotics and theory of mind).

Later did work on squirrel-like robot deception: received much more positive feedback!

Next studied *deception via dishonesty*. Birds and other animals may feign strength/death in order to convince the predator that they’re more intimidating than they are.

! Peacocks: carry a heavy tail around, but it’s useful for attracting other mates. It says “I’m so fit but you can’t do anything.” Human analogue: rolex! (vs. fake rolex, vs. everyone wearing a rolex).

Study: investigated the risk associated with investing resources in appearing stronger than you are (like a peacock carrying its heavy tail).

! Finding: some criteria, in some mob sizes, these strategies work quite well.

Further studies: deception by omission, deception by commission (actively sending a signal vs. not saying anything).

! Finding: robots can use a reduction of frustration in pill sorting tasks to keep elders engaged.

Ethical Issues:

Q: Should a robot be allowed to lie? Use deceptive methods? A: Exemplar ethical frameworks in considering robotic deception (violates categorical imperative). Might be more utilitarian, ensure stakeholders benefit.

Q: Should we be making deceptive robots? How does one ensure it is only used in appropriate context

Q: Is there an inherent right whereby humans should not be lied to?

IEEE established deception guidelines: “It is necessary to develop recommendations regarding the acceptability of deception...”. Put forward guidelines for incorporating deception.

.....

3.4 Covey Award Talk: John Weckert on Our Technology Fetish

Russell said roughly “nothing is so silly that philosophers can’t talk about it”.

3.4.1 Background

We have to admit: humans aren't physically great at anything relative to the rest of the natural world (swim, fly, run, walk, climb, digging). Even our sensors aren't great.

! We're good at thinking.

So: We need technology to survive! ! We are technological creatures

“[Technology is] the improvement brought about on nature by man for the satisfaction of his necessities...therefore the main aim of technology is to promote good life, well-being, and adapting the medium to the will of the individual” (Ortago y Gasset)

So, two points: 1) We need technology to survive, 2) We need it not just for survival, but for leading a good and flourishing life.¹

But:

Lots of current skepticism of science and experts. “We should be concerned about the growing neglect of science and about the overt attacks on it from different sides” (Coenen 2019).

Examples: anti-vaxers, climate change skeptics. Big concerns! Science is well established.

But, less skepticism of technology. Tends to be an *overwhelming* love of technology.

We have a fetish with technology: fetish an excessive and irrational devotion or commitment to a particular thing.

3.4.2 Do we have a tech fetish?

Now onto the central question;

Central Q: Do we have an excessive and irrational devotion or commitment to technology?

Our worldview: 1) technological progress is *good*, 2) new technology is good even if not real progress.

! Something like a Kuhnian paradigm: technological progress is accepted without question.

Techno-optimism: “life is getting better. Poverty continues nose diving; adult literacy is at an all-time high; people around the world are living longer, living in democracies, and are better educated than at any other time in history...” (Will Reinhart, 2018)

Techno-pessimism: “Without a responsible and steady guiding hand, it become useless, and perhaps detrimental. Humanity must recognize the limits of technology and look to more realistic solutions for modern problems” (Sebastian Miller, the dangers of techno optimism ,2017).

¹Some might argue technology is more about the latter.

! But, there is still a fetish! “While many people feel fearful about the increased use of AI, machines, and robots in the workplace the move to a more automated future is **inevitable**.”

O’Regan consulting: 20% of top 1000 US businesses plan to roll out AI methods *this year*. Highlights the inevitability.

Critical: *development continues in spite of concerns*. Why? Well: 1) benefits outweigh risks, 2) we have a technological fetish, 3) technological fixes preferred over life style changes***Dave: Yeah!.

! We are on a technological treadmill. We bring in new tech, raises new challenges, we need new tech for the challenges, rinse and repeat.

Levels of thought about technology:

1. Philosophical: good life, flourishing, awareness of problems
2. Development of new technologies; innovation, new products.
(The fetish is primarily at this level: “develop it! it’s good! keep doing it.)
3. Use: regulations, policy.

It seems: excessive and irrational devotion to computer technology. People call out for: 1) faster, more memory, 2) quantum computers, 3) Moore’s – why does it matter if it ends? we have good computers now! 4) Networks, 5) AI.

! These developments are seen as *good in themselves!*

Example: state surveillance (see: China). China has made a music video promoting the importance of integrity and trustworthiness ahead of the national rollout of the controversial Social Credit System. This can measure and enhance “trust’ nationwide and to build a culture of sincerity.

! “It will forge a public opinion environment where keeping trust is glorious. It will strengthen sincerity in government firms, social systems, and so on ...”.

Takeaway: perhaps this isn’t the ideal world we want to live in.

3.4.3 Quotes on What Computers Should Do

Q: What shouldn’t computers do?

A: In the past, Weizenbaum, Moor, Kuflik, and Lenman have given arguments; the future/present, Bostrom, Hawking, and Musk.

Wiezenbaum²:

²Creator of Eliza

1. “The asking of the question ‘what does a judge know that we cannot tell a computer?’ is a monstrous obscenity. That it has to be put into print at all, even for the purposes of exposing its morbidity, is a sign of the madness of our times... the point is that computers should not be given such tasks.” (1970s).
2. “All projects that propose to substitute a computer system for a human function that involves interpersonal respect, understanding, and love, are obscene. Their very contemplation ought to give rise to feelings of disgust in every civilized person.” (1976)

Q: What is the problem?

Weiz A: Lots of things! 1) Suggests humans are machines, 2) suggests machines are better than humans, 3) reduces human contact, 4) reduces scope for care, 5) reduces ability to develop virtues, 6) It’s dehumanizing.

James Moor: “when computers improve life, its the moral thing to do to use them...the values on which computer decisions are made must always be a human responsibility.”

Arthur Kuflik: “Even when decision making responsibility is given to computers, humans should be kept in the loop.”

James Lenman: focuses on finding meaning in life and flourishing – computers can be used to these ends.

Attitudes toward the future of AI:

More recently, Nick Bostrom: superintelligence represents an existential risk to humanity.

Hawking and Musk: don’t let AI take our jobs or kill us.

Floridi: logically possible but implausible.

Boden: Dream on. AGI is a pipe dream.

McKinglyay: robots won’t kill us in our lifetime.

“If-Then” style argument: if something is possible in the future, we should do something about it now. Nordmann rejects this premise.

Q: Should we worry about future problems?

A: Perhaps computers will never be intelligent! Floridi says computers can be smart but not necessarily conscious.

! The case hasn’t been made that we shouldn’t worry about future problems. Really, it’s that we shouldn’t ignore them (but need to also focus on pressing current issues).

3.4.4 Returning to the Fetish

Returning to the fetish: most philosophers work at the level of the philosophical; what is the good life? How does technology/AI contribute to our flourishing? But the fetish is at the second level; develop new technology.

Overriding values in the world: productivity and efficiency! Make tech that makes things more productive.

The fetish: “the culture of consumption is one mostly created by the concept of planned obsolescence, whereby the products are intentionally designed to last only for a short period and then you will go buy a new one.”

In 2016: estimated waste of 50 million tonnes from throwing out old phones/computers. [Dave](#): :O.

This fetish does matter: what sort of world do we want to live? There’s a fundamental disconnect between how philosophers are thinking about the impacts of this tech on the world, but people develop them without thinking deeply about them.

Q: Should the precautionary principle be considered?

A: In some places, the answer is obviously no (the USA), whereas others, the answer is obviously yes (Europe).

Some lingering questions:

Science and technology are guided to some extent by the world’s needs.

Q: Can we do anything? ! Funding determines research priorities. ! industry focuses on certain kinds of projects (usually that benefit the company’s bottom line).

Good example: medical technology. Used to be a rule of thumb that 90% of medical research was focused on diseases that pertain to only 10% of the world’s population (those with money).

Q: Should we do anything? ! A: Largely a question of values! ! Sir Robert Watson said there must be a transformative change to human civilization if we are to avoid an extinction crisis. By transformative change, we mean a fundamental, system-wide reorganization across technological, economical, and social factors”

Should we do anything?

Raises questions about death. Epicureans attitude toward death: we don’t worry about having not been here 100 years ago, so why do we worry about death?

! Common view is that it’s bad for the dead person to be dead. That’s nonsense!

We have an obsession with longevity! As opposed to flourishing. A long bad life is probably worse than short epic one.

Raises other questions: How should we get enjoyment in life? Should we control the population? How should we be thinking about our relationship with non human nature?

Conclusions:

We absolutely have a technology fetish.

What kind of world do we want to live in? ! Transfer autonomy to machines? ! Live in a surveillance society?

Are we losing the benefits of being in nature? Are we getting too far from how we evolved?

Questions we need to think about!

If we're going to come to grips with this obsession, we have to look at some of these big issues to do with how we want the world to look (and not just the technical ones).

.....

3.4.5 Erlant Etxeberria on Mathematical Explanations and Counterpossibles

Focus: mathematical explanations in physical sciences.

Example: Cicada bugs! Some of them come out every year, some come out every n years, where n varies from region to region. Their life cycle (when they “emerge”) follows an unusual period: how do we explain this phenomena?

! Spend most of their life underground, but every 13 or 17 years they emerge.

Life cycle of the 13 year cicada: they emerge as full adults, climb up into a branch of a tree, and lay eggs in the tree. The newborn cicadas are above ground, fall from the tree, and spend the rest of their lives (up until laying eggs) underground!

Q: Why? This is odd!

A: Many predators have 2-5 year life cycles. So, by cycling by a large prime number of years, cicadas avoid their predators. 13 and 17 year cycles cannot be tracked by any smaller number.

**This is a mathematical explanation of an empirical phenomenon.

Different kinds of mathematical explanations: 1) within mathematics (intra-mathematical); some theorems such as $a^2 + b^2 = c^2$ have *explanatory* proofs, 2) Explanations of empirical phenomena that use math, like physical laws, 3) Extra-mathematical distinctively mathematical explanations of empirical phenomena.

The puzzle here is the *third* one: how does this take place?

Some answers:

1. Counterfactual accounts of causal explanations: Woodward (2003) developed an interventionist account of causal explanation that relies on counterfactual conditionals: if A had been different, then B would have been different, too (if A explains B).
2. Causal and non-causal explanations: Peter Lipton: “why, in a class of 22 students, there are four students whose birthday falls on the same day of the week? The answer is that there are seven days in a week, 22 students, so any arrangement makes this the case.”
 - ! Can the counterfactual account be extended to cover non-causal explanations?
 - ! One idea: Had there been 21 students, there needn’t be 4 birthdays the same day of the week.

Definition 10 (The Problem of Counterpossibles): *When counterfactuals require an antecedent that is contrary to mathematical facts.*

! Had 22/7 = 3, there needn’t be 4 birthdays the same day of the week! So, this isn’t well grounded.

Solution from Baron, Colyvan, and Ripley (2017): If, in addition to 13 and 1, 13 had factors 2 and 6, then North American periodical cicadas would not have 13-year life cycles.

Worries: Can’t even conceive of what it would be like for math to be different in this way.

! Answer: impossible things can be entertained! Water could be H₃O, and so on. Sometimes all we need is to state something to entertain the idea?

Problem: impossible worlds! “Possible worlds, I think, we should be willing to accept, at least if they are given a sober actualist explanation. There are possibilities, different ways things could have been or could still be. But impossible worlds, impossibilities, ways things could not have been, that is too much to swallow” (Stalnaker 1996)

Answer: The BCR theory is not concerned with the semantics of counterfactuals or how to evaluate them, but instead in a more general project.

Real worry: avoiding contradictions!

BCR’s strategy: avoid needless contradictions when dealing with counterpossible antecedents. Three main notions in all counterfactual reasoning: 1) holding fixed, 2) twiddling, 3) ramifying (all indicating how much is being changed).

Back to Cicada example: counterfactual that explains why cicadas have 13 year life cycles. We need to change something about 13: originally we said that perhaps $2 \cdot 6 = 13$, so 13 is no longer prime (but this yields a contradiction).

! Response: avoid, rather than manage, contradictions. Idea is to introduce the right kind of change to the conditions that BCR calls ‘ α ’ a new operator: $2 \cdot \alpha 6 = 13$. Then, BCR holds under

Q: Problem? Yes! We get what they wanted in the first step (no immediate contradiction), but contradictions still easily follow.

A possible response: analogy with regular counterfactuals. So: the rock shattered the window because, if the rock hadn't been thrown, the window wouldn't have shattered.

! But this creates an inconsistency! The rock *was* thrown.

Problem: The analogy breaks: counterfactuals of any kind requires that we deal with an impossible antecedent. Lewis (1979) gives us some criteria to assess similarity of worlds based on match of histories and break of laws or miracles.

! In math, a change in the structure is *permanent*. It has to propagate to the rest of mathematics.

So: what are we left with? We can't just say 13 isn't prime any more.

Open questions for dealing with the puzzle:

1. Deny existence of genuine non causal explanations of empirical phenomena
2. Accept mathematical explanations, but analyze them in non-counterfactual terms
3. A better approach to counterfactuals that doesn't require impossible worlds.

.....

3.5 Anders Herlitz on Predictive Fairness

Main point: impossibility theorem from Kleinberg that should have significant implications outside of CS.

Background:

COMPAS: a statistical method that was used by US courts to determine recidivism risk (re-offense rate). Used this risk score to prosecute, set bail, sentencing and so on. ! Algorithmically generated risk score had major influence on outcomes.

COMPAS systematically ascribed higher risk scores to black nonrecidivists compared to the risk scores ascribed to white nonrecidivists.

Northpointe, the company that developed COMPAS replied: COMPAS is not biased, it's well calibrated. When it identified a set of people who had probability p of being recidivists, a p -fraction of this set were recidivists. So: it's fair!

Thus, two notions of fairness:

1. If it systematically ascribes higher scores to negative instances in one salient social group compared to another, and/or ascribes lower scores to positive instances in one group vs. the other.

2. Predictive decision method is unfair if it systematically ascribes higher risk scores to one salient social group compared to another.

Problem: no imperfect decision system can meet both conditions! (Result from Kleinberg).

Proof idea: first, need some definitions.

Definition 11 (Well calibrated): *An algorithm identifies a set of people as having probability z of constituting positive instances then approximately a z fraction of this set should be indeed positive. This condition should hold when applied separately in each group as well.*

Balance:

1. Balance for positive class: avg score received by people constituting positive instances should be same in both groups.
2. Balance for negative class: avg score received by people constituting negative instances should be same in both groups.

Impossibility argument: notation. N_t is number of people in group t , P_t number of people in t who are positive, x_t be avg score given to a member of group t who is negative, y_t same but positive.

Proof idea:

Supposed the decision method is well-calibrated: then, the total score ascribed to a group can be written as a linear function:

$$(N - P_t)x_t + P_t y_t = P_t.$$

If the decision method is balanced, x and y must be the same for each group. So, defines a system of equations.

Solving this system of equations requires that one of two conditions hold: either P is the same for each group (so all x and y satisfy every equation), or $x = 0$ and $y = 1$.

Q: Why care? Sure, COMPAS seems wrong, but doesn't that mean we shouldn't rely on decision algorithms in the judicial system?

A: Hold on... The impossibility argument shows that *all systematic methods* that carry out this probabilistic reasoning are doomed. This applies to all areas where we make predictions: judicial system, healthcare, and so on (whether it is algorithmically- or human-driven).

Normative suggestion: think about principles of fairness for predictive decision methods. Here are three:

1. *Dominance*: A decision method that is better w.r.t. one of the dimensions of fairness and worse with respect to non of the dimensions is better overall.
2. *Transparency*: Unfair decision makers should be aware of unintended differences in impact.

3. *Priority to the worse-off* : Decision method should aim to be better for members of a worse-off subgroup.

.....

4 Friday June 7th

The last day!

4.1 Alan Wagner on Kantian One Day, Consequentialist the Next

Main Idea: Often faced with situations where you need to trade-off between different ethical frameworks.

Example: walking past a homeless person on the street. How do the particulars of that day effect whether you give aid to the individual or not? What about if the person is injured?

! Previous psych study: participants were asked to attend an ethical symposium, along the way come across someone injured/in need. If the participant is in a rush, they tend not to help.

The Problem: There is no agreed upon correct moral theory, and decisions vary by user and context.

Robotics community: tend to want robots that act ethically.

**Their approach: generate action recommendations for a robot based on several different ethical frameworks.

Method: survey different populations to see what they think are different ethical actions (ask both average adults and ethics experts)

Three different ethical frameworks: 1) utilitarian framework focused on consequences, 2) Kantian framework focused on ethical duty, 3) social justice framework (fairness across groups), or 4) virtue ethics.

! Focus on cases where these frameworks *disagree*.

! Also, approach includes *moral emotions* like shame, pride, empathy, and so on. These emotions might effect whether a choice is “right”.

Q: So, what should a robot do? Let’s say in the giving-aid-on-the-street case.

A: Form an action selection arbiter that takes as input suggestions for different ethical frameworks *and* contextualized model of human moral emotions (guilt, shame, pride, and so on). Pictured in Figure 1.

Experiment 1: Under which circumstances should one allow the opponent to win? Factors to consider: emotional state (is a child frustrated?), number of games played, age, others.

Experiment 2: Pill sorting for elder care. Should the robot provide certain kinds of feedback? Say, false feedback that still encourages the person?

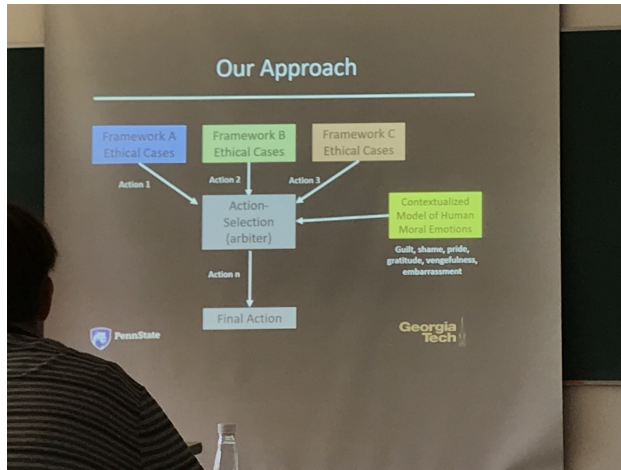


Figure 1: Approach for ethical action arbitration and decision making.

! Survey for ethical “answers”: what do average everyday people say to do in these contexts (use AMT to collect data). Also collect “expert opinions” from ethicists.

Contextualized Moral Emotions: robot retains model of emotional states of human participants; the models affect action selection.

Compare three action selection strategies:

1. Pill sorting: compare 1) always truthful (even when harsh), 2) Ethical expert recommendations, 3) Folk morality recommendations.
2. Game playing with children: 1) play to win, 2) ethical expert, 3) folk morality.

Conclusions:

Robot may need to use different ethical frameworks at different times.

These frameworks may not generate same action recommendations.

Robot may be able to use moral emotions to select best framework.

Possible trade-off between expert and folk opinions.

.....

4.2 Angelica Fleury on Creating Moral Ideals in Robots from Buddhism

Goal: Introduce Buddhist principles for automated moral machines and contrast them from existing moral decision making methods.

Q: Can robots make moral decisions?

A: Amanda Sharkey in “Can we program or train robots to do good?”: she says no! Searle also would say no (via chinese room): strong AI can’t understand their decisions, so can’t make a moral decision.

Against Robot-Morality Skeptics:

Require an assumption about a metaphysical “self” tied to understanding. Necessary components for understanding morality.

Humans have a self, and this is typically sought after in robot design.

We ought to build a moral machine as nearly human as possible, requires claims that humans are the moral standard.

Objections; need for inner self as a basis for agent to have capacity fo be moral might be mistaken! Also, we should take into consideration disputes on the need for a metaphysical self. ! Buddhists challenge the ethics of the self as something to be sought after.

Q: Should humans serve as a moral standard?

A: Might be the right starting point to model moral machines after, but not the standard. Our own agenda allows us to act selfishly (according to most ethical frameworks).

Objective: Not to create a metaphysical moral criteria outside human influence, but instead to progress towards understanding how to make a moral machine.

Q: What are the necessary components for a Buddhist-influenced AMA?

A: (from Danstini, Floor, Meyer 2006): 1) Autonomy (minimal sense, not Kant), 2) Emotions (need to be reciprocal, guide expression and motivation), and 3) Emotion regulation (maintain desirable emotions).

Approach 1: program agents with emotions via BDI-Based agent programming language.

Approach 2: program agents with emotions via ALMA (a layered model of affect).

! Suggestions for modifying agents with emotions: 1) Program initial mood to maximize empathy, 2) Appraisal (less reactionary, motivated by compassion), and 3) Coping with artifical emotion (their still triggered by if-then statements, human understandable).

Example: Patient with dementia who is in need of assistance, but providers are reluctant to interact with patient due to emotional distress. Provider may neglect necessary interactions, whereas a moral machine can take into account negative emotions and recognize the need for interruption and its negative emotions can be overridden.

Conclusions:

Don’t get too hung up on metaphysics of self

Reconsider that moral machines need to be just like humans

Keep emotions around but use mindfulness as a model to incorporate emotion.

.....

4.3 Simon Award Talk: Juan M Durán on Philosophically Analyzing Simulations

Focus: Come up with a systematic approach for studying computer simulations from a philosophical perspective (concentrating on two viewpoints).

Outline:

1. First viewpoint: Problem-solving technique viewpoint
2. Second viewpoint: Description of patterns of behaviour viewpoint: 1) historical definitions (1960-2013), 2) common topics, 3) philosophical assumptions.
3. Philosophical implications/studies of computer simulations.

4.3.1 View 1: Problem Solving Techniques (PST)

Idea: Computer simulations are about solving a problem.

Definition 12 (PST (take one)): *(Teichroew and Lubin 1966) \Simulation problems are characterized by being mathematically intractable and having resisted solution by analytic methods"*

Definition 13 (PST (take two)): *(Humphrey 1990): \A computer simulation is any computer implemented method for exploring the properties of mathematical models where analytic methods are unavailable."*

Common Themes from PST definitions: 1) Simulations are about finding the set of solutions to mathematical models, 2) Use of simulations is justified when analytic methods are unavailable, and 3) exploring *properties* of mathematical models.

Three key assumptions (for when to use computational simulations according to PST):

1. Analytic intractability
 - ! Simulations are a tool for computing. Instrumental role for finding results for a model. Understood as epistemically inferior to analytic methods.
2. Implementation *simpliciter*
 - ! Assumed that mathematical models can be directly implemented as a computer simulation. No genuine methodology mediates between the models and simulation.
 - ! Simulation models are ontologically on par with mathematical models.

3. Inherited representational capacity

/ Simulations inherit their representational capacity of a target system from the mathematical model they implement.

Recent paper from Bueno and Colyvan (2014) high *representation* as a critical piece of theories of computer simulations.

4.3.2 View 2: Description of Patterns of Behavior (DPB)

Again, a historical record;

Definition 14 (DPB (take one)): (Shubik 1960): *“a simulation of a system is operation of a model or simulator which is the representation of a system*

Definition 15 (DPB (take two)): (Bristwile 1980) *“Simulation is a technique for representing a dynamic system by a model in order to gain information about the underlying system”.*

Definition 16 (DPB (take three)): (Shannon 1998): *“The process of designing a model of a real system and conducting experiments with this model for the purpose of understanding the behavior of the system and evaluating various strategies for the operation of the system.”*

Main idea: simulations are units of analysis in their own right (in the same way that scientific tests are).

Commonalities across these definitions: 1) Computer simulations are primarily concerned with descriptions/representations of the target system, and 2) Simulations allow inferences about properties of the target system.

Three key assumptions (for when to use computational simulations according to DPB):

1. A proper methodology for computer simulations.
/ Computer simulations become units of analysis in their own right.
2. Own/proper representational capacity of the simulation.
3. New forms of knowledge.
/ Capable of giving rise to new knowledge that we couldn't access through other means.

4.3.3 Philosophical Implications

Q: What can we learn about scientific explanation from these views of computer simulation?

A (Traditional view): mathematical model can give explanation for real-world phenomena, but not necessarily computer simulation (view from Kros 2008, Weirich 2011).

! Scientific explanation: “Simulation model does not provide an acceptable explanation of the material system” (Krohs 2008).

Explanation for the PST:

Analytic Interpretability ! “Theoretical models need to be further analyzed or solved to provide descriptions of and predictions about the dynamics of the world they model.” (Krohs).

Implementation *simpliciter*

Inherited representational capacity ! Since the simulation does not have explanatory input on its own and since it is limited to relating the theoretical model with the real world phenomena, the simulation must inherit the representational capacity of the mathematical model.

Q: Do computer simulations have more to say about explanation? Or is that it? (the PST story)

A: Yes, explanation is better understood through the DPB

** Better view (from the DPB): When we explain, we don’t explain the real world, *we explain the results of a simulation* (and hope the simulation is linked to the real world).

! To solve this explanatory difficulty, we can draw a distinction between *knowledge* and *understanding*. Explanations can give us *understanding*, but not knowledge.

Key Idea: results actually shed light on the simulation itself; they help us *understand* the simulation.

Explanation for the DPB:

Scientific explanation: “It is the simulation model, and not an exogenous mathematical model, the unity with the most explanatory relevance for the results of the simulation” (Durán 2017).

A property methodology for computer simulations ! Simulations must be units of analysis in their own right.

Self representational capacity of the simulation ! It makes possible the identification of what can be ascribed to the world and what is an artifact of the simulation

New forms of knowledge ! computer simulations have a **genuine explanatory role** and thus ground their epistemic power.

4.3.4 Conclusions

Recap: two distinct viewpoints on computer simulations and their role in scientific explanation.

! PST and DPB are two different theoretical frameworks about the philosophical study of simulation. PST strongly depends on mathematical models, where DPB view suggest a new methodology and epistemology of computer simulations.

Important note: not all definitions qualify as PST or DPB (though these two comprise most of the literature). Some others: 1) Computer simulations are arguments (Beisbart 2012), 2) Computer simulations as unfolding the content of a well-described scenario (El Skaf, Imbert 2013).

Conclusion: if we can come up with well established views that unify perspectives/definitions, we can take ourselves further in this literature.

.....

4.4 Alan Wagner on The Aristotelian Robot

Focus: Philosophical treatment of a what it would look like to construct an Aristotelean robot.

! Guiding Q: How can we construct a robot that can make use of virtue ethics?

! Virtue ethics focuses on the agent, while deontology focuses on the maxim that orients moral action, and utilitarian theories focus on outcomes.

Central Claim: Aristotelian ethics is an optimal framework to implement for an artificial social agent. Reasons:

It can be implemented as a moral learning framework

It can take into account moral psychological and the role of moral emotions.

Q: Are humans by nature moral?

A: No! We *learn* to be educated by our community and our experiences. we are “Aristotelian” humans first, later we may become Kantian, utilitarians, or nihilists. At first we are taught virtues using moral narratives and engaging their moral emotions (moral pedagogy through narration can help teach us about morality!). *Dave: I remember reading the book of virtues as a kid!*

Q: How did humans become moral?

A: societal pressures and evolution generated the conditions for morality. Moral emotions played a critical role.

***Morality is a *societal accomplishment*.

Agents can learn to be moral by interacting with a community: virtues are embedded in the community.

Aristotle claim: core virtues are *quasi-universal*, and should appear in most groups: 1) self-respect, 2) gratitude, 3) courage. When a *con ict* arises among exemplars, core virtues are engaged.

! Moral conflict can cause reflection/introspection: why is something important? Social conventions/norms unburdened us from constant moral reflection.

Q: How do we train a moral robot?

Approach: use moral narratives to teach!

1. Traditionally moral training takes the form of narratives: Odysseus, Mulan, the Bodhisattva in the Jatakas.
2. Consider the hypotheticals: “What should X do in this situation?”.
3. Challenge: using narratives in this way requires emotional understanding!

Aristotelian Framework:

Questions Considered:

1. Are we intrinsically moral? ! No
2. Moral pedagogy ! Aristotelian model
3. Does reason move us to morality? ! No
4. Can we be moral without a sense of moral persona? ! No
5. Did we make ourselves moral? ! Yes
6. Does the moral “I” depend on moral narratives? ! Yes
7. Does reason depend on affect to tell us what is moral? ! Yes
8. Feelings required for moral emotions
9. What’s the role of phenomenology in morality?

Conclusion:

Should be post-Cartesian and post-Turinian; we need to connect mind and body and recognize the importance of narrativity in morality.

Just having embodiment is not enough.

Cannot separate emotions from embodiment.

.....

4.5 Himavath Jois on Should Robots be Allowed to Punish Us?

Scenario: an elementary school sets up a mixed robot/student team to build a bridge. In a typical team, if a student misbehaves, the teacher might scold them (put them in time out, etc.). But in this joint robot-student team, should a robot be allowed to put a misbehaving student in timeout?

Guiding Q: Should robots be allowed to punish people in some circumstances?

“Humans tend to cede decision making authority to machines in mixed robot-human teams” (Gombolary et al.).

Q: What is punishment for? Why is it necessary?

A: Punishment is critical for teaching! (See: education, sports, norms)—especially when mistakes are costly (military).

Q: Will humans accept punishment from a robot? And in particular, how will people respond?

Naive Hypothesis: People will prefer to be punished by a robot (not advocating for this, but speculating), even when the punishment is physical!

Experiment: explore whether this hypothesis has merit. Setup involves a punishment device: a robotic exoskeleton capable of physical restriction as punishment for a wearer’s action.

! Key ethical frameworks: consequentialism, conserving autonomy, utilitarianism.

Consequentialist view: if the punishment leads to decreased wrongdoing, we might view it as a good thing (but, we have to consider the decrease in autonomy/pleasure of the individual)
! Toward autonomy: Bentham’s Panopticon. Too much surveillance decreases autonomy. Would we prefer emotional control or physical control? Or just object to this loss of autonomy?

Utilitarianism vs. Retributivism: Goal is to administer a punishment that decreases crime/wrongdoing. Retributive would be eye-for-an-eye.

Three scenarios:

1. Scenario 1: A new house arrest device allows prisoners to live outside of prison but limits what they can do.
2. Scenario 2: An army uses exoskeletons to maintain control over recently surrendered soldiers.
3. Scenario 3: A world leader uses robotic exoskeletons to control a civil population and limit unrest.

In light of these scenarios, conduct experiments that vary the type of punishment on individuals. Can receive any of: 1) no punishment, 2) verbal scolding, 3) physical restriction (through the exoskeleton).

! Contrast a human initiator vs. a robotic initiator (for the punishment).

Preliminary Results:

Pilot subjects presented with Godspeed survey: ! reported that physical punishment was very frustrating ! subjects that underwent robot initiated punishment found the exoskeleton more intelligent than the human

Minimal understanding of subject preferences

Motivating subject to trade accuracy for speed is difficult. ! Currently in the process of revisiting the methodology ! So this is the only beginning (but again: we're not advocating for this to happen). Just want to initiate the academic conversation, better understand human preferences about these topics.

Q: Why should we even think about this? A: "The machine does not isolate man from the great problems of nature but plunges him more deeply into them", Antoine de Saint-Exupery in *The Little Prince*.

.....

4.6 Presidential Address: Don Berkich on Underdetermination after Computation

Q: should the philosophy of science community catch up to new computational methods? Do philosophers of science care about these advances? (they should!)

Definition 17 (Naive View of Science): *A theory proposes (in the strong deductive sense), an experiment disposes*

One try: Observation \Rightarrow : Hypothesis. But that can't be right. "Observation" isn't propositional.

Note: really want to understand what we mean by "observation" in the computational age.

Second try: ExperimentalObservation \Rightarrow : Hypothesis. Where, "ExperimentalObservation" is the proposition that "specific states of affairs incompatible with the truth of hypothesis have been experimentally observed to obtain."

Q: Why is this naive view of science naive?

A1: ExperimentalObservation is "theory laden" (see: Duhem 1906 and Quine's Two Dogmas 1951). No putatively observational statement is observational.

A2: If ExpObs is theory laden, then falsification is ambiguous.

A3: If falsification is ambiguous, then scientific theory is underdetermined by the evidence for it (see: Duhemian Holism or Quinean Holism).

A4: Is scientific theory is underdetermined by the evidence for it, then there are no “crucial” experiments in science.

A5: If no crucial experiments, then theory choice is determined by extra-scientific values.

) : Theory choice is driven by extra-scientific values.

Q: What are the extra-scientific values in question?

Duhem A: Experimenter must display “good sense” in choosing theory.

Quine A: Scientists may appeal to such principles like minimal mutilation that tend to protext math and logic.

! Upshot: By good sense/minimal mutilation/other values, naive view of science is naive insofar as it ignores the necessary role of these values in theory choice.

Holism: hypotheses are affected by: Science *and* math *and* Logic *and* Computability.

Q: How does the widespread deployment in the service of science affect Duhem and Quine’s accounts?

Examples of computational methods used in science (on the side of the entailing proposition in the naive view): hurricane modeling, cellular modeling, neural modeling, ecological systems, and so on.

! Computer modeling/simulations are just one half of the computational turn in the sciences. We also have data collection.

*These pieces are so powerful, we should update the view of science to: ComputationalObservation \Rightarrow : Hypothesis, where CompObs is the proposition that “specific states of affairs incompatible with the truth of Hypothesis have been algorithmically inferred from vast data-troves automatically collected and classified from multiplicities of sensors.

Q: What is the observation when perceptual experience is no part of CompObs?

Discussion: Is CS epistemically privileged? Thinking about Quine’s implicative view, is CS distinct from math/logic/science in some sense?

! View 1: CS is science, enjoying no greater privilege than the rest of science. ! View 2: CS is like math/logic, epistemically privileged either by being *a priori* or defensible as occupying a central commanding position in our web of belief.

Conclusion: Wanted to ask: how does CS factor into our scientific theory? Is it epistemically privileged in some sense? Each possibility visited represents trouble. Moreover, computationall mediated observation may turn out to be a shaky foundation on which to build the edifice of science (as we’re doing).

Dave: And that's a wrap!

.....

References

- [1] Richard W Byrne and Andrew Whiten. Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans. 1990.
- [2] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69 (345):118–121, 1974.
- [3] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.
- [4] Edward Stein. *Without good reason: The rationality debate in philosophy and cognitive science*. Clarendon Press, 1996.
- [5] Peter M Todd and Gerd Gigerenzer. *Ecological rationality: Intelligence in the world*. OUP USA, 2012.
- [6] Karen Yeung. hypernudge: Big data as a mode of regulation by design. *Information, Communication & Society*, 20(1):118–136, 2017.