Comment on this doc, send me a note at james@cham.net, or @jamescham on Twitter.

Situated AI (bit.ly/situated-ai)

The best summary of the state of AI ethics is Brian Christian's new book The Alignment Problem. He spent four years talking to the leading researchers and practitioners thinking about AI right now, summarizing their insights and making ties between the limitations and implications of AI models. The book highlights limitations of the current AI ethics conversation in his last chapter. I think it points to a bigger limitation to the discussion.

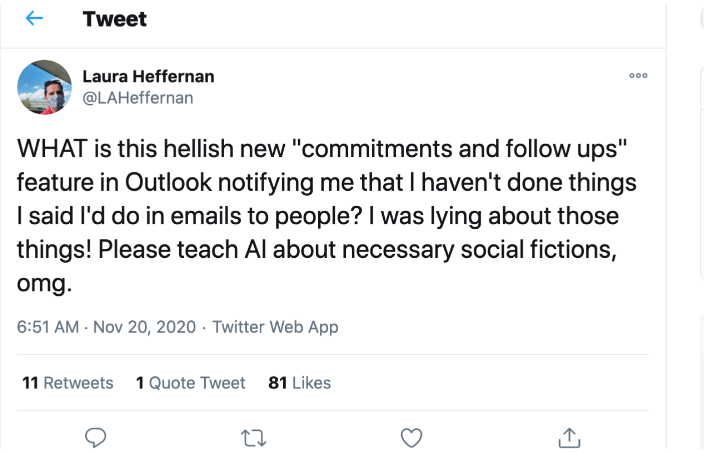
The limitation is that we take AI models too seriously, treat humans as static unchanging abstractions, and think of the algorithms as the center. But AI models only exist in relation to humans. We build, inform, enforce, ignore, obey, and manipulate them. Those interactions are the actual center of the most interesting AI ethics challenges.

A few thinkers are talking about this interaction--taking humans seriously and also understanding the strengths and limitations of AI models.

Here's a running list. I'm looking for other examples!

- **Training AIs to work under "constitutions":** Constitutional AI: Harmlessness from AI Feedback. "As AI systems become more capable, we would like to enlist their help to supervise other AIs. We experiment with methods for training a harmless AI assistant through self improvement, without any human labels identifying harmful outputs. The only human oversight is provided through a list of rules or principles, and so we refer to the method as 'Constitutional AI'".
- **Designing AIs like Juries:** Jury Learning: Integrating Dissenting Voices into Machine Learning Models "We introduce jury learning, a supervised ML approach that resolves these disagreements explicitly through the metaphor of a jury: defining which people or groups, in what proportion, determine the classifier's prediction."
- **How AIs get legitimacy:** How Humans Judge Machines (summary). Forthcoming book by economics professor Cesar Hidalgo. A very clever study on human moral intuitions around AI. When do humans blame machines more or less? You can think of this as a "moral census" of American attitudes towards AI. Presents opportunities for when AI models have more or less legitimacy. Will change over time, and in addition to comparing human vs machines, would be interesting to add how humans judge organizations.

- **AI should be viewed as a bureaucracy rather than God-like judge:** Street Level Algorithms (summary). Ali Alkhatib and Michael Bernstein. CS paper that deserves more attention. Looking at how laws are actually enforced, and corrected by people on the ground, and comparing that with the way that AI models are actually enforced. AI models are often blamed for being wrong but they are actually easier to correct than laws, and when corrected, they are centralized.
- **People cheat and fool AIs (social science view):** Seeing Like a Finite State Machine. Henry Farrell. Smart blog post (and reference to James C Scott's Seeing Like a State) by a professor of International Relations. "There is a very plausible set of mechanisms under which machine learning and related techniques may turn out to be a disaster for authoritarianism, reinforcing its weaknesses rather than its strengths, by increasing its tendency to bad decision making, and reducing further the possibility of negative feedback that could help correct against errors." AI models are profoundly conservative: they reflect the world that we tell them about. And in that way, left unchecked, they can be manipulated and given inaccurate information.
- **Who gets to decide?** The Allocation of Decision Authority to Human and Artificial Intelligence. Joshua Gans, Susan Athey. The principal trades off an AI's more aligned choice with the need to motivate the human agent to expend effort in learning choice payoffs. When agent effort is desired, it is shown that the principal is more likely to give that agent decision authority, reduce investment in AI reliability and adopt an AI that may be biased. Organizational design considerations are likely to impact on how AI's are trained.
- **Hybrid intelligence:** Hybrid Intelligence: A Paradigm for More Responsible Practice. "The envisioned practice would harness human and machine complementarities to develop systems of human-machine hybrid intelligence. Such systems integrate the best capabilities of both machine intelligence and human users while mitigating the deficits of each. In this approach, outcomes can be improved not only by improving the underlying technologies but also by improving the human-machine collaboration processes."
- **People cheat and fool AIs (CS view):** Ethical Algorithms, Chapter 3. Michael Kearns, Aaron Roth. "The lesson that when an app is mediating or coordinating the preferences of its users (as opposed to simply using their data for some other purpose, such as building a predictive model), the algorithm design must specifically take into consideration how users might react to its recommendations—including trying to manipulate, defect, or cheat."
- **Economics of AI:** Prediction Machines. Ajay Agrawal, Avi Goldfarb, Josh Gans. Popular business book by 3 economists. AI models aren't necessarily better. But they are much, much cheaper and consistent than the human-based judgements.
- **People power models:** Ghost Work. Mary L. Gray and Siddharth Suri. By looking closely at the people who are providing the judgments used to build AI models and taking them seriously, you can see ways to improve those judgements and see the boundaries of the data being captured.

- The State Machine. Yudhanjaya Wijeratne. Short fiction on an AI run world.
- Adapting to the algorithm. Kyle Chayka. "As I write more about how algorithms influence culture, I find myself using the word "adaptation" quite often. I analyze how creators and content have to "adapt to the algorithm" in order to find an audience on a platform."
- A Conversation about Human Computation. Walter Lasecki. "Two things are exciting to me. First is the ability to look at how the traditional organizational structure intersects with how we think about human computation at the task level — that is, incorporating both the micro and macro structures into the design of "hybrid intelligence" systems. The second thing is real-time applications of human computation methods — thinking about ways to create new systems and tools to provide human-in-the-loop answers within seconds of the data becoming available."
- Her: You should probably add the film "Her"2013 as a reference to human - AI interface. It's contributed to my vision for http://timeOS.co and steered my deep thinking about the potential of a Personal OS. @rmishiev via Twitter.
- High Tech Modernism Henry Farrell and Marion Fourcade. "Instead, we claim that algorithms are much better understood as a manifestation of hierarchy than as an extension of market competition."
- A New Breed
- The Perfect Match, Ken Liu. "You thought Centillion was just an algorithm, a machine. But now you know that it's built by people—people like me, people like you." Ajay Agrawal
- 

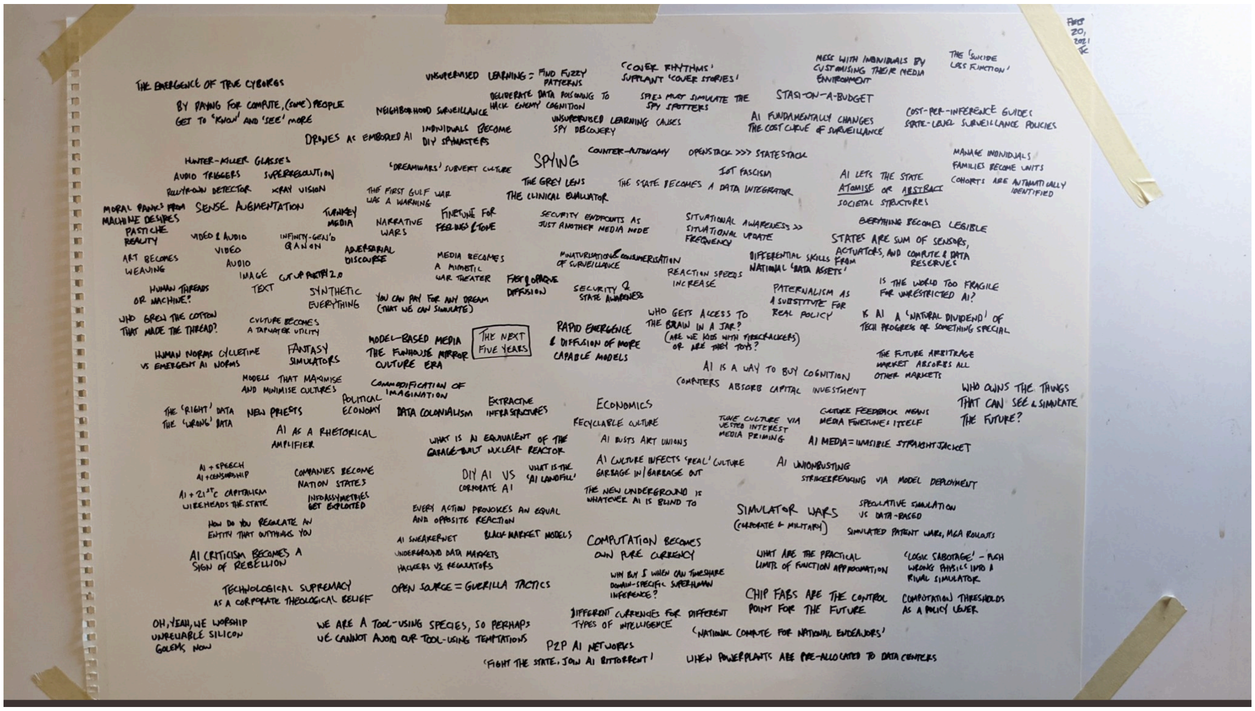
  Laura Heffernan
  @LAHeffernan

  WHAT is this hellish new "commitments and follow ups" feature in Outlook notifying me that I haven't done things I said I'd do in emails to people? I was lying about those things! Please teach AI about necessary social fictions, omg.

  6:51 AM · Nov 20, 2020 · Twitter Web App

  11 Retweets 1 Quote Tweet 81 Likes
-

- https://twitter.com/jackclarksf/status/1363238616365899780?s=21


THE EMERGENCE OF TRUE CYBORGS
BY PAYING FOR COMPUTE, (SOME) PEOPLE GET TO 'KNOW' AND 'SEE' MORE
DRONES AS EMBODIED AI
HUNTER-KILLER GLASSES
AUDIO TRIGGERS
SUPERRESOLUTION
RULLYROWN DETECTOR
XRAY VISION
MORAL PANICS FROM MACHINE DESIRES
SENSE AUGMENTATION
PASTICHE REALITY
VIDEO & AUDIO
VIDEO
AUDIO
IMAGE
TEXT
ART BECOMES WEAVING
INFINITY-GEN'D QANON
CUT UP POETRY 2.0
HUMAN THREADS OR MACHINE?
SYNTHETIC EVERYTHING
WHO GREW THE COTTON THAT MADE THE THREAD?
CULTURE BECOMES A TAPWATER UTILITY
HUMAN NORMS CYCLETIME VS EMERGENT AI NORMS
FANTASY SIMULATORS
MODELS THAT MAXIMISE AND MINIMISE CULTURES
THE 'RIGHT' DATA
THE 'WRONG' DATA
NEW PRIESTS
AI AS A RHETORICAL AMPLIFIER
UNSUPERVISED LEARNING = FIND FUZZY PATTERNS
NEIGHBORHOOD SURVEILLANCE
DELIBERATE DATA POISONING TO HACK ENEMY COGNITION
INDIVIDUALS BECOME DIY SPYMASTERS
UNSUPERVISED LEARNING CAUSES SPY DISCOVERY
'DREAMWARS' SUBVERT CULTURE
SPYING
THE GREY LENS
THE CLINICAL EVALUATOR
THE FIRST GULF WAR WAS A WARNING
TURNKEY MEDIA
NARRATIVE WARS
FINETUNE FOR FEELINGS & TONE
ADVERSARIAL DISCOURSE
MEDIA BECOMES A MIMETIC WAR THEATER
FAST & OPAQUE DIFFUSION
YOU CAN PAY FOR ANY DREAM (THAT WE CAN SIMULATE)
MODEL-BASED MEDIA
THE FUNHOUSE MIRROR CULTURE ERA
THE NEXT FIVE YEARS
COMMODIFICATION OF IMAGINATION
POLITICAL ECONOMY
DATA COLONIALISM
EXTRACTIVE INFRASTRUCTURES
'COVER RHYTHMS' SUPPLANT 'COVER STORIES'
SPIES MUST SIMULATE THE SPY SPOTTERS
COUNTER-AUTONOMY
OPENSTACK >>> STATESTACK
IOT FASCISM
THE STATE BECOMES A DATA INTEGRATOR
SECURITY ENDPOINTS AS JUST ANOTHER MEDIA NODE
SITUATIONAL AWARENESS >> SITUATIONAL UPDATE FREQUENCY
MINATURISATION & CONSUMERISATION OF SURVEILLANCE
REACTION SPEEDS INCREASE
SECURITY & STATE AWARENESS
RAPID EMERGENCE & DIFFUSION OF MORE CAPABLE MODELS
WHO GETS ACCESS TO THE BRAIN IN A JAR? (ARE WE KIDS WITH FIRECRACKERS) OR ARE THEY TOYS?
AI IS A WAY TO BUY COGNITION
COMPUTERS ABSORB CAPITAL INVESTMENT
ECONOMICS
RECYCLABLE CULTURE
AI BUSTS ART UNIONS
TUNE CULTURE VIA VESTED INTEREST MEDIA PRIMING
MESS WITH INDIVIDUALS BY CUSTOMISING THEIR MEDIA ENVIRONMENT
THE 'SUICIDE LOSS FUNCTION'
FEB 20, 2021 JC
STASI-ON-A-BUDGET
AI FUNDAMENTALLY CHANGES THE COST CURVE OF SURVEILLANCE
COST-PER-INFERENCE GUIDES STATE-LEVEL SURVEILLANCE POLICIES
MANAGE INDIVIDUALS
FAMILIES BECOME UNITS
COHORTS ARE AUTOMATICALLY IDENTIFIED
AI LETS THE STATE ATOMISE OR ABSTRACT SOCIETAL STRUCTURES
EVERYTHING BECOMES LEGIBLE
STATES ARE SUM OF SENSORS, ACTUATORS, AND COMPUTE & DATA RESERVES
DIFFERENTIAL SKILLS FROM NATIONAL 'DATA ASSETS'
PATERNALISM AS A SUBSTITUTE FOR REAL POLICY
IS THE WORLD TOO FRAGILE FOR UNRESTRICTED AI?
IS AI A 'NATURAL DIVIDEND' OF TECH PROGRESS OR SOMETHING SPECIAL
THE FUTURE ARBITRAGE MARKET ABSORBS ALL OTHER MARKETS
WHO OWNS THE THINGS THAT CAN SEE & SIMULATE THE FUTURE?
CULTURE FEEDBACK MEANS MEDIA FINETUNES ITSELF
AI MEDIA = INVISIBLE STRAIGHTJACKET
AI + SPEECH
AI + CENSORSHIP
AI + 21ST C CAPITALISM WIREHEADS THE STATE
COMPANIES BECOME NATION STATES
INFOASSYMETRIES GET EXPLOITED
HOW DO YOU REGULATE AN ENTITY THAT OUTTHINKS YOU
AI CRITICISM BECOMES A SIGN OF REBELLION
TECHNOLOGICAL SUPREMACY AS A CORPORATE THEOLOGICAL BELIEF
OH, YEAH, WE WORSHIP UNRELIABLE SILICON GOLEMS NOW
WHAT IS AI EQUIVALENT OF THE GARAGE-BUILT NUCLEAR REACTOR
DIY AI VS CORPORATE AI
WHAT IS THE 'AI LANDFILL'
EVERY ACTION PROVOKES AN EQUAL AND OPPOSITE REACTION
AI SNEAKERNET
BLACK MARKET MODELS
UNDERGROUND DATA MARKETS
HACKERS VS REGULATORS
OPEN SOURCE = GUERILLA TACTICS
WE ARE A TOOL-USING SPECIES, SO PERHAPS WE CANNOT AVOID OUR TOOL-USING TEMPTATIONS
AI CULTURE INFECTS 'REAL' CULTURE
GARBAGE IN / GARBAGE OUT
THE NEW UNDERGROUND IS WHATEVER AI IS BLIND TO
COMPUTATION BECOMES OWN PURE CURRENCY
WHY BUY $ WHEN CAN TIMESHARE DOMAIN-SPECIFIC SUPERHUMAN INFERENCE?
DIFFERENT CURRENCIES FOR DIFFERENT TYPES OF INTELLIGENCE
P2P AI NETWORKS
'FIGHT THE STATE, JOIN AI BITTORRENT'
AI UNIONBUSTING
STRIKEBREAKING VIA MODEL DEPLOYMENT
SIMULATOR WARS (CORPORATE & MILITARY)
SPECULATIVE SIMULATION VS DATA-BASED
SIMULATED PATENT WARS, M&A ROLLOUTS
WHAT ARE THE PRACTICAL LIMITS OF FUNCTION APPROXIMATION
'LOGIC SABOTAGE' - PUSH WRONG PHYSICS INTO A RIVAL SIMULATOR
CHIP FABS ARE THE CONTROL POINT FOR THE FUTURE
COMPUTATION THRESHOLDS AS A POLICY LEVER
'NATIONAL COMPUTE FOR NATIONAL ENDEAVORS'
WHEN POWERPLANTS ARE PRE-ALLOCATED TO DATA CENTERS

To Read:

- **The Future of Work in the Age of AI: Displacement or Risk-Shifting?** "We contend that economic forecasts of massive AI-induced job loss are of limited practical utility, as they tend to focus solely on technical aspects of task execution, while neglecting broader contextual inquiry about the social components of work, organizational structures, and cross-industry effects...we highlight four mechanisms through which firms are beginning to use AI-driven tools to reallocate risks from themselves to workers: algorithmic scheduling, task redefinition, loss and fraud prediction, and incentivization of productivity." TH
- Seeing Like a State – James C. Scott talking about how governments make human life "legible" by imposing formalisms and a kind of literal as well as figurative terraforming; i.e., you need a fixed address in order to be found by the tax authorities. ML is doing something like this in the digital sphere. BC
- An Engine, Not a Camera – This looks at financial models and how the models sometimes become in a sense "realer" than the phenomena they were designed to model. BC
- What to Expect When You're Expecting Robots – MIT roboticist Julie Shah, in addition to being a childhood friend, is one of the leading experts on human–robot collaboration (e.g., in aerospace manufacturing). An interesting look at humans and ML systems working in this case literally elbow-to-elbow. BC
- https://ali-alkhatib.com/blog/digital-forests. AA
- https://twitter.com/infoxiao: AI mediated trust. AA.
- Computer supported, cooperative work conference. AA

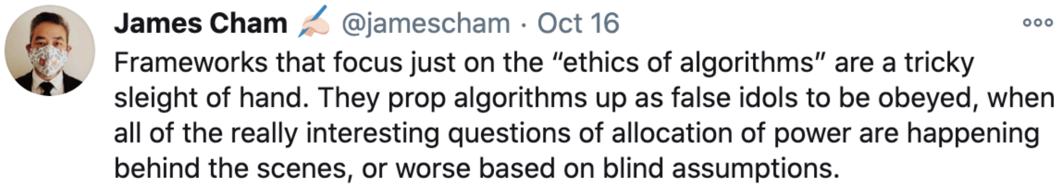
- Automating inequality, Virginia Eubanks. AA
- Manipulation-Proof Machine Learning "But when consequential decisions are encoded in rules, individuals may strategically alter their behavior to achieve desired outcomes." DR
- Causal Strategic Linear Regression "In many predictive decision-making scenarios, such as credit scoring and academic testing, a decision-maker must construct a model that accounts for agents' propensity to 'game' the decision rule by changing their features so as to receive better decisions." DR
- **New Laws of Robotics**—Frank Pasquale. "We now have the means to channel technologies of automation, rather than being captured or transformed by them." FP
- Bandwidth by Eliot Peper. "Like everyone else, Dag relies on his digital feed for everything—a feed that is as personal as it is pervasive, and may not be as private as it seems. As he struggles to make sense of the dark forces closing in on him, he discovers that activists are hijacking the feed to manipulate markets and governments."
- Critiquing Algorithms: a list. Anonymous.
- Fauxtomation and hidden labor in technology: a list. Anonymous.

TZ suggestions: Here are a few papers that came to mind based off of our chat earlier today:

- Formalization as a Social Project, Philip Agre (it's an oldie, but a goodie for thinking about blindspots even in the current ML paradigm)
- Towards a Rigorous Science of Interpretable Machine Learning, Finale Doshi-Valez & Been Kim (in-depth look at interpretability and its limits)
- Hard Choice in AI: Addressing Normative Uncertainty through Sociotechnical Commitments, Roel Dobbe, Thomas Gilbert, Yonatan Mintz (philosophical, risk framing in AI systems)
- Accountability of AI Under the Law: The Role of Explanation, Doshi-Valez et al. (limits of AI from legal perspective)

What is this field called?

- Situated AI
- Human-AI Interface
- Other names?



James Cham ✍🏻 @jamescham · Oct 16

Frameworks that focus just on the "ethics of algorithms" are a tricky sleight of hand. They prop algorithms up as false idols to be obeyed, when all of the really interesting questions of allocation of power are happening behind the scenes, or worse based on blind assumptions.

I'm looking for other examples of this--and challenges!

Comment on this doc, send me a note at james@cham.net, or @jamescham on Twitter.