Eight Views Reviewed:
A Review of the Literature of Why AI Matters Morally

Bryan A. Knowles
Information School
Educational Policy Studies
UW-Madison
baknowles@wisc.edu

Alan Rubel
Information School
Center for Law, Justice, and Society
UW-Madison
arubel@wisc.edu

November 2019

Abstract

In this exhaustive (read, "til convergence") review of the AI Ethics literature, I identify eight theoretical views presented here. Each theoretical view is characterized by its own recurring themes, rhetorical moves, priorities, assumptions, and/or disciplinary thinking. For each theoretical view I synthesize the literature operating under that view and critique them towards the question, "Why does AI matter morally?" Since no one theory can be used to summarize all eight perspectives, and since I resist ontologizing these views, I conclude with an eye towards my own future work and UW-Madison's goals of teaching AI Ethics under CDIS. Appendices attend to terms and abbreviations, methods, and limitations.

Example footnote[1] *Example endnote*[(i)]

Keywords: *Artificial Intelligence, Ethics, Literature Review*

Contents

[1] Example footnote

1 Introduction

Before I ask, "How should we teach AI Ethics," I should answer, "What is being said about AI as a moral matter?" AI is a relatively recent term in the literature, and policy matters regarding AI are far from settled. AI clearly matters. Just recently a self-driving Uber struck and killed a jay-walking pedestrian because it's AI was not trained for jay-walkers: the AI did not believe her to be human and so it did not stop. Pedestrian safety is not the only moral matter for AI, and the conceptions for how we begin to think about these matters are varied.

Therefore, it is my goal in this literature to enumerate those conceptions. This is an exhaustive (til convergence) review of theories (conceptions) of AI as a moral matter.[(ii)] My goal is to identify the central concerns in this area and to organize my results into conceptual headings: nine headings, one excluded here which will be the standalone subject of a future paper, leaving my "eight views reviewed." I attempt a neutral representation, resisting early ontologization in favor of letting themes emerge from the discourse itself; and I attempt to write towards an audience of general scholars who may have an interest in AI, AI Ethics, AI Policy, or so on. To balance the wide scope of this review with the needs of my audience, I rely on lay terms as much as possible and provide propositions above and summaries below each subsection in the body.

See my appendix on Methods for more information on how I searched, read, and wrote.

What comes out of this review is a "canvas" of conceptions on AI in the literature. In the initial search, I came across several related canvases and key canvases repeatedly cited by others. Each is has its own goals, and each has goals different from my own.

So, I give Table 1 below to put these into context with one another and with my own review. Table 1 is a *very* high-level overview of the canvases in the literature. Many of these will be referenced in more depth in specific (sub-)sections below.

Source *Annotation*	**Major Headings/Divisions**
Russel, Stuart and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson. 2009. *Textbook recommended by Burton et al, includes in one section this list of ethical concerns unique to or exacerbated by AI*	• Loss of Jobs • Excess/Scarcity of Leisure • Existentialism • Illuse • Accountability Gap • Human Extinction
Gunkel, David J. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT. 2012. *Concerned with what rights a machine should have in a society*	• Moral Agency • Moral Patiency • Other Approaches
Mittelstadt, Brent, et al. "The Ethics of Algorithms: Mapping the Debate." *Big Data and Society*. 2016. *A map of the AI Ethics debate connecting each of the three general concerns to the issues exacerbated by AI: Unjustified Actions, Opacity, Bias, Discrimination, Challenges to Autonomy, Challenges to Privacy, Responsibility Gaps*	• Epistemic Concerns • Normative Concerns • Traceability (Concern)
Muller, Vincent C. *Fundamental Issues of Artificial Intelligence*. Springer. 2016. *A collection on AI and its impacts on philosophy*	• Computing • Information • Cognition and Reasoning • Embodied Cognition • Ethics
Burton, Emanuelle, et al. "Ethical Considerations in Artificial Intelligence Courses." *arXiv Preprints*. 2017. *Themes guiding how to teach Ethics in an introductory AI course, connected to three common philosophical frameworks (deontology, utilitarianism, virtue ethics) and applied to case studies: Robots as Care Givers with Competing Objectives, Human Trust in AI Systems, Bias, Reliability, Oversight, Assessment, Distribution of Benefits, Weapons*	• AI Behavior • Supply of Jobs • Killer Robots • Singularity • Moral Standing
Petit, Nicolas. "Law and Regulation of Artificial Intelligence and Robots: Conceptual Framework and Normative Implications." 2017. *A map of AI Ethics literature along the first two posts, recommending taking the stance of the third, choosing which of the first two to employ for Discrete Public Interests, Systemic Externalities, Existential Issues*	• Ex Post/Legalistic • Ex Ante/Technological • Benevolent City Planner
Calo, Ryan. "Artificial Intelligence Policy: A Primer and Roadmap." 2017. *A primer on AI Policy on specific concerns in these areas*	• Justice and Equity • Use of Force • Safety and Certification • Privacy and Power • Taxation and Displacement of Labor • Others
Van Wynsberghe, Aimee and Scott Robbins. "Critiquing the Reasons for Making Artificial Moral Agents." *Science and Engineering Ethics* 3(25). 2018. *A critique of reasons given by Machine Ethicists for studying and creating machines with moral reasoning built-in* . Poulsen, Adam, et al. "Responses to a Critique of Artificial Moral Agents." *arXiv Preprints*. 2019. *Responses to the above*	• Inevitability • Prevent Human Harm • Complexity Requires Moral Subroutine • Better Public Trust • Prevents Immoral Use • Better Morality than Humans • Better Morality in Humans

Continued from previous page	
Raso, Filippo, et al. "Artificial Intelligence & Human Rights: Opportunities & Risks." *Berkman Klein Center*. 2018. *A risk assessment looking at each of these sectors along the axes of different rights: Freedom from Discrimination, Equality before the Law, Life/Liberty/Security, Freedom from Arbitrary Arrest/Detention/Exile, Fair Public Hearing, Innocent until Proven Guilty, Privacy, Freedom of Opinion, Peaceful Assembly/Association, Desirable Work, Adequate Standard of Living, Education*	• Criminal Justice • Finance • Healthcare • Content Moderation • Human Resources • Education
Samaniego, Jose Miguel. "Practices of Governing and Making Artificial Intelligences." *Disputatio* 7(8). 2018. *A map of the processes of AI standardization, from STS theory, along these scales, finding Themes, Group Formations, Entities and Actors, Logics, Domains, Relationships to Standards, Ontologies, Standardization as World-Making Process*	• Statements • Literatures • Actors and Networks • Cosmoses
Rahwen, Iyad, et al. "Machine Behavior." *Nature* 568. 2019. *A "machine behavior" framework applied to these issues and modeled on ecology recommending two views of questions, Proximate and Evolutionary; two views of the objects of study, Static and Dynamic; and five views of the scale of inquiry, Individual Machine, Collection of Machines, Humans shaping Machines, Machines shaping Humans, Human/Machine Co-Agency*	• News Ranking • Algorithmic Injustice • Self-Driving Vehicles • LAWS • Algorithmic Trading • Algorithmic Pricing • Online Dating • Conversational Bots

Table 1: High-level overview of AI Ethics and adjacent canvases in the literature, in chronological order

In most, the goal is give a *taxonomy of AI issues and concerns*. These taxonomies are either organized by grouping up into hierarchies or given as a flat list.

Raso et al instead gives a taxonomy of the *sectors* where the effect of AI has been closely watched in recent years, presented alongside an enumeration of rights affected, for better or worse, in these applications (2018). Rahwen et al (2019) and Petit (2017) both give analytical frameworks for thinking about these emerging AI issues. And Samaniego gives a detailed look across different scales of complexity where the *raising* of these concerns is put into practice; he does this in order to critique the relationships between the act of raising AI Ethics concerns and the actors who benefit from that act (2018).

My (sub-)sections are meant to be a *taxonomy of conceptualizations* of AI Ethics: eight *views* reviewed. I developed my headings from concept maps, and I developed these concept maps by drawing out the arguments and motions in each piece in the review, so my headings should be seen as an attempt to group up and explain AI concerns by how writers *reason* about them in practice. Jumping from one (sub-)section to another, for example, the rhetorical moves I critique differ. I want to stress *rhetorical moves* instead of the more

specific term *argument*: the authors in this space are not… always precise. Where a relevant argument is given that adds something new to the review, I critique it; generally though, my subject of inquiry is the "discursive formation"[iii] of AI as a moral matter. Most accurately, this is a "canvas of conceptions."

A summary to show how each section fits onto the canvas follows:

Professionalism Here I attend to the view that answers "Why does AI Ethics matter" with "because it's our job" and at the same time explores the idea of the "job" from a Sociology lens. The Professionalism view is the black sheep of the group, but I use this perspective to set the stage for the others because, as a developer and teacher of developers, I am not excluded from it; and as developers and/or teachers of developers, you my readers are not excluded from this perspective either. It *is* our job, this is clear, but saying that teaches little.

Public Trust Here I attend to the view that begins from the assumption that AI can only be successfully deployed if the public has reason to trust it and that reason has been conveyed well to them.

Moral Machines Here I attend to the view most closely aligned with the Machine Ethics literature. Machine Ethics tries to create artificial moral agents and/or studies under which conditions this could actually be achieved. I do want to note that the discursive formation of this section is larger than what was caught in this review (with one brief exception[iv]): this is the view that asks first, what *is* a moral machine? The Machine Ethics literature then proceeds to ask how it could be done; science fiction, on the other hand, asks that question alongside asking, what is a *human*? Machine Ethics and science fiction owe a lot to one another, and I know of one teacher[v] who uses the latter to get at the former.

Democracy and Societal Effects Here I attend to another view that looks at AI from the scale of the public. Public Trust originally grew out from this as a special category, but was split off in a later draft. Democracy and Societal Effects primarily deals with concerns about negative effects of AI on *procedures* center to the maintenance of a democratic society. This is divided into the special interests in representation by data, mediation of discourse, and scrutiny in specific institutions.

Jobs Here I attend to the long list of economic sectors AI has been projected to affect and the question of what a government is to do to mitigate displacement issues that will inevitably follow.

Consumers Here I attend to the other side of the same coin as Jobs, where the view begins with real examples of how consumers have been affected, usually negatively, by AI. This is divided into visible and invisible concerns of AI/human interaction.

Policy and Guidelines Here I attend to an overview of the legal perspective on AI Ethics. This is divided into the who, the why, and the how: who is making AI policy; what are the general goals of that policy; and what are the general legal frameworks one could apply.

Personhood And here I attend to a special question grown out of Policies and Guidelines on what moral standing an AI should have in society and the legal implications of our different choices.

Epistemic Trust My goal is to attend to this view in a standalone paper. I do not feel I can do this view justice in the space here. This is the view that asks, when looking at an AI-generated research paper, if that paper's "results" could be trusted. How we draw the line matters when, for example, an AI's "testimony" is trusted over that of a human in court.[(vi)]

2 Professionalism

2.1 AI matters morally because AI is a matter of current professional debate; the work of this profession has societal consequences; professional groups are a building block of society; debating AI Ethics contributes to shaping the identity of the professional group; and one's sense of that identity contributes to shaping their mundane ethics

Two Theories on Professional Identity Speaking generally (not about AI), professional identity matters morally. Professional identity, in Durkheim's view, is like a horizontal backbone to society, a complement to the vertical hierarchies of self, family, city, state, nation, and so on. Durkheim goes on to say that Western society cannot operate without professional identity because Western society *uses* that identity as a source of rule:

> Political society as a whole, or the state, clearly cannot discharge this function [determining the rules for division of labor, how to resolve conflicts related to this, and how to maintain peace and order related to this]. Economic life, because it is very special and is daily becoming increasingly specialised, lies outside their authority [central government] and sphere of action. Activity within a profession can only be effectively regulated through a group close enough to that profession to be thoroughly cognisant of how it functions, capable of perceiving all its needs and following every fluctuation in them. The sole group that meets these conditions is that constituted by all those working in the same industry, assembled together and organised in a single body. This is what is termed a corporation, or professional group (1985).

Laws and group dynamics both play a part in regulating an individual's behavior. For we programmers, AI lies at the heart of the stories we tell about our profession. AI is inherited from Alan Turing's challenge to create an AI intelligent to imitate a human, and this challenge is given to the field's newcomers. It is in these stories that we determine who we are and construct our professional identity. And we programmers, it could be any symbol at the heart of our identity, but it *is* AI, if only for today. It is towards that goal that so much money and energy is spent.

But this is an uncritical view of Durkheim. Where do these professional structures come from? Is their maintenance just and fair? Are their borders necessary? Has it really

served us in a complementary way to state rule? And of AI, our understanding of it has grown considerably since Turing's challenge, both technically and socially. How ethical could a human-level machine be that is built on the poor working conditions of Amazon Turk workers exposed to graphic images of child pornography, in service of training an automaton to recognize these images for us? This is not the paper to answer these questions, but raising them points towards the kinds of issues that this view on AI Ethics cares about.

Kenneth Burke gives another theory on professional identity; Burke tells us that vague membership criteria in a professional, scientific, or social group may be used as a rhetorical appeal to discriminate against the would-bes entering the field, or to delegitimize the opinions of those having already entered (1969). Burke is writing after the second World War, Durkheim before the US officially entered the first, and Burke's subject matter is figuring what exactly happened to give power to Hitler's rhetoric. He, like Durkheim, centered on identity, but here he sees it as a phenomena of rhetoric. We are at birth separated, but it is through identification that we connect our experiences with one another.(vii) And here professional identity is not a property of the *structure* of society that maintains us, but instead a something that is appealed to by individuals in order to achieve the orator's[2] ends.

Comparison to Engineering Is AI a profession? AI, Computer Science, and Data Science, these are similar to other professions, like Engineering and Health, because all five touch the public broadly and infrastructurally and can lead to wide harm when they fail; but only Engineering and Health, in these examples, are "professionalized." Not everyone can dress up like a doctor, but anyone can learn to code.

Open source software movements show another goal of Computer Science: to tear down its borders. And this includes recent, publicly-released AI toolkits that enable laypersons to adapt modern advancements in technology for their own contexts. The story goes in the professional identity that it is dangerous to centralize powerful tools, even under well-

[2]I mean orator more generally, as anyone who invokes some appeal of interest, whether in writing, speech, dance, website design, etc.

intentioned stakeholders.

Another story is that software is like a bridge. Engineers have been studying how to build bridges for millennia and have near perfected the process of planning, deploying, and maintaining bridges.[viii] Software is something constructed too–it's just that we programmers have spent much less time at it. And because ethics matters in bridges, it's an easy segue to say ethics also matters in software.

Burton et al (*supra* Out) extend this to talk about the duties of we *teachers of* programmers (2017). Like Engineering schools, we have a duty to ensure public safety by training and gatekeeping new recruits into the profession. "As instructors," they say, "we want to develop curriculum that not only prepares students to be artificial intelligence practitioners, but also to understand the moral, ethical, and philosophical impacts that artificial intelligence will have on society."[ix]

Codes of Ethics Codes of Ethics matter morally because by creating codes (an activity involving many conversations and many actors) we put into words, for ourselves, our professional identity. (I will talk more about Codes of Ethics from a Policy standpoint later.) The stories we tell, the goals we share, and the sense and procedures of our professional identity, these have consequences for general society because we members of the profession fall back on these in making everyday moral decisions. And, thinking hard about these goals determines how we want to move forward as a team; this can either strengthen our professional identity or fragments us into complementary parties. AI, if not *the* thing of debate, is still of high substantial interest.

Codes of Ethics are not the only normative force here: Garzcarek and Steuer compare codes to company guidelines and policies where the developer is employed; institutional oaths sworn by the developer; codes already adopted by scientific societies and agreed to by the developer as part of their membership in the society; and religion (2019). Any of these can shape the moral behavior of the individual, though too much regulation might be

stifling. An opponent of adopting new Codes of Ethics[x] might point to any one of these as overlapping and would therefore have any of many reasons to claim Codes as redundant and therefore unnecessary. Still, Garzcarek and Steuer argue that Codes *create* the profession. Codes, oaths, uniforms,[xi] flags, rituals, and so on, these give those within the profession a sense of belonging.

So, Garzcarek and Steuer look at example Codes from ACM (representing Computer Science), ASA (Statistics), and GI (Informatics), and they work through the common concerns brought up in these Codes. Codes spell out the priorities of our profession, they argue; Codes spell out what "counts" as our profession and what counts as *work in* our profession. And the formal processes of deliberating and updating those Codes, occurring at regular intervals, those authors argue, keep the Codes socially relevant. Revision proceedings permit the members of the group to redefine for themselves what matters for their professional identity. "[T]he morality of the data science community," they conclude, "is evolving and…it is a shared task to develop it." Data science is not *yet* a community, their argument requires. Coming together to create a new Codes under the name of a new field, *despite* possible redundancy with existing policies and so on, this uniquely serves to *name* our new field as a field. And this will be a field for which ethical responsibility, *as a goal*, will come from within. So, is AI a profession? We get to decide.

Summary 2.1 The scholarly space of this section holds that professional groups are building blocks of modern society, and even if AI is not a professional group of its own, it is a substantial topic of professional debate. Our work, we programmers, has societal consequences. And our identity as members of our profession does guide us, as in making major design decisions, or in practitioners carrying out day-to-day, mundane ethics.[xii] And even if our professional group proclaims open borders, there is *some* barrier to entry one cannot deny; those barriers may be falling with new AI toolkits becoming publicly available, but I cannot in good faith believe that those resources will outright replace highly specialized AI

researchers: most toolkits require graduate-level math. And lastly, even if Codes are redundant or fail to bring out their tenets, the endeavoring itself is still necessary in a society like ours: an endeavor that demands accountability, sense of professional identity, and autonomy from the State.

3 Public Trust

3.1 AI matters morally because AI can benefit the public; the public must feel comfortable in order to switch to AI from previous non-AI alternatives; and policy made that deploys AI without the public's trust violates that comfortability

Trust After Failure Public AI projects sometimes fail. One argument is that, as a consequence, members of the public will lose trust in the AI, the system the AI was embedded in, and/or policymakers involved with those systems.

Margetts and Dorobantu (*supra* Dem) write in part about just such failures (2019). These failures can stem, for example, from inaccurate datasets used to train the AI or from AI models that were not tested for bias. And the context of the failure matters, Margetts and Dorobantu argue: if it is a bad Netflix recommendation, then I am likely to be very forgiving; but if it is a public project, carried out by police (as Margetts and Dorobantu are looking at) on uninformed visitors of a carnival, and an arrest is made by faulty facial recognition and out of date warrant information, then, well, the bar there is higher. Public AI projects also require *some* check on the application of the AI: for example, Margetts and Dorobantu tell of public school officials who have no idea how they should act in the face of AI-reported probabilities about their students. Public trust in the *AI* might not matter: the public can lose trust in the policymakers and their understanding of the social contexts they are introducing AI into.

Margetts and Dorobantu give the short recommendation to include the public in government-sponsored AI project decisions through "citizens' juries." In these, members of the public are invited to rank their preferences towards explainability and accuracy, or so on. "Trans-

parency," they say, "is crucial for assuring public trust." And including members of the public in the design loop goes a long ways towards that.

Trust in Different Cultures Members of the public will have to make a switch to AI from existing non-AI alternatives; and to respect their autonomy, they must feel comfortable with the "moral (software) code," so to say, of the AI. However, as one study finds, what "feel comfortable with moral code" means likely varies by culture.

Awad et al argue, talking about self-driving vehicles, that in order for a public to adopt and accept that kind of AI technology, the public must be able to trace and trust the origin of the vehicle's moral code (2018). For self-driving vehicles to become widespread, consumers[(xiii)] will need to feel comfortable to make the switch. Consumers and pedestrians[(xiv)] alike, as self-driving cars occupy more and more of public roadways, must trust that this autonomous technology will be safe and moral.

Awad et al ask, how do people across the world rank Trolley scenarios differently? For example, are residents of Florida more likely to rank animals over humans in a choice of who the self-driving car should spare if forced to choose who to kill? Collecting data through their MIT Moral Machine experiment, they find three top level cultural clusters: Western, Eastern, and Southern (Latin America). Although the "winner" in each comparison category (old vs. young, man vs. woman, etc.) is nearly consistent across countries, they find that *relative frequencies* across categories allows responses to be grouped by similarity. These groupings give rise to their cultural clusters, and this leads them to argue that AI moral preference must be fine-tuned to the culture(s) it is deployed in: when particular human/AI values are misaligned, the public may not understand, or may disagree with, AI's moral "code."

Why Trust *AI* Differently? One argument is that, as far as assessment and policy are concerned, AI technology is like existing non-AI technology in many ways; but, the argument continues, it is different in that AI has a *vast* range of applications and AI's

emergent consequences (once deployed within a network of other technologies, including other AIs) are not well understood.

The CDT, responding to the High-Level Expert Group's (HLEG) *Draft Ethics Guidelines for Trustworthy AI*, claim that "[i]n many contexts and applications, truly trustworthy AI remains hypothetical" and that trust is not all that matters (2019). The HLEG should also focus, the CDT argues, on *continuous assessment* of the *entire socio-technical* system an AI is deployed in. And when looking at "trust," one should look for its necessary components–an ethical purpose, a robust implementation, and governance over the socio-technical system–, and one should shift from thinking about "trust in *technology*" to trust "in the *systems* within which the technology is used and the processes that govern its use."

They go on to recommend the HLEG these nine questions, to be used when assessing how trustworthy an AI system is:

1. "Do people have a non-AI alternative. . . ?"
2. "Have the effects of the risk mitigation or management plan been tested?"
3. "Has there been an assessment of the potential to improve the risk mitigation or management measures?"
4. "Have the effects of interactions between multiple AI systems been identified?"
5. "Is/are the system(s) being used as intended?"
6. "What measures have been taken to limit unintended uses?"
7. "What impacts might the system have on the fundamental rights of the intended users?"
8. "What impacts might the system have outside of the intended group?"
9. "When systems make decisions impacting people other than the users. . . are the criteria for balancing risks and benefits to the public communicated?"

AI has almost nothing to do with this.

Eight of these nine questions could be asked of any technology system. And this makes sense given the CDT's focus on the system outside the AI: this system *will* include older tech-

nologies, laws, norms, tech/human interactions, etc. So the CDT provides *these* questions, and they provide them *now*, because others *aren't already* asking them. Technology systems in general have at least some certain moral responsibilities that must be met. For example, all public-effecting technology systems should apply equally to members of the public, be understood by the public, not violate an individual's right to choice, and be appropriately distributed over sub-populations (questions 1, 7, 8, 9). And all public-effecting technology systems should not impose undue risk or burden on members of the public and should not be used by individuals or public officials to undermine members' rights (questions 2, 3, 5, 6, 7, 8).

But question 4 stands out.

AI differs from previous technology in two ways: AI has a *vast* reach in application, and humans do not understand the networked, automated interactions that could arise from multi-AI systems. These emergent[3] patterns could conceivably widen the accountability gap (*supra* Pol) without any mechanism of control. Question 4, as I imagine it, is an attempt to check under the bed for the AI bogeyman, emergent behavior. The bogeyman of the 90s may have been the threat of a worldwide computer virus making its way uncontrollably around the internet and destroying now-connected financial systems, sending the world economy into chaos. The fear here is much the same, but somehow even less tangible: there is no virus circulating, and yet still the world plunges into chaos, as something that emerges, "just happens," unpredictably from network interactions of AI agents independently operating. *Ex post*[4] accountability be damned when the stakes are that high. So, as the CDT recommends, the only feasible alternative to holding AI systems accountable so we can trust in them is through continuous and systemic assessment.

[3]Formally, as defined in Complex Systems Science.

[4]*Ex post* accountability is accountability for wrongs after they have occurred. *Ex ante* accountability is accountability in prevention of a would-be wrong. Compare laws and policies that apply after a car crash has occurred to those mandating that we turn our headlights on at night.

Summary 3.1 This scholarly space suggests that AI morally should benefit the public. The public must feel comfortable with AI, as they will be required to make a switch from previous non-AI alternatives. And policy morally should not violate this comfortability. Yet, while public trust is fundamental, it should not be left to public outcry to hold public uses of AI accountable, and *ex post* accountability is not justifiable when the stakes are high. So, the only feasible measure (of those suggested so far in this review) for ensuring public trust in high stakes public uses of AI is *ex ante* continuous and systemic assessments.[(xv)]

4 Moral Machines

4.1 AI matters morally because there is pressure to create the Artificial Moral Agent; creating AMAs may improve our own moral reasoning; AMAs test our intuition about human judgement before- and during-the-fact; AMAs test our intuition about true dilemmas, finding creative alternatives to them, and avoiding them altogether; AMAs test our intuition about trusting non-humans to make moral decisions in human best interest; and AMAs test our intuition about how we forgive mistakes and evaluate tasks, for humans and machines

Is making AMAs morally worth it? One argument is that having AMAs at our service will cause humans to lose our skill of (or become worse at) making moral decisions for ourselves. But a rebutting argument is that *the process of making* AMAs will make humans better at moral reasoning because of the hard work to get there.

Van Wynsberghe and Robbins critique, in part, that the argument (made elsewhere) that machines can be better moral agents than humans is a hard argument to accept because it requires a known moral truth which humans already do not have access to or agreement on (2018). Still, even when they entertain that argument, they argue that "[o]utsourcing our moral reasoning to machines could cause a moral deskilling in human beings."

Erica Neely (in Poulsen et al 2019[(xvi)]) rebuts. She first agrees along similar lines that an AMA[5] will likely not be much different than an ethicist in the way that an ethicist is not *that* much better (not better by computational leaps and bounds) than a layperson at moral

[5]Machine with moral reasoning built-in, loosely speaking, an "Artificial Moral Agent."

thinking. And she agrees that machines may be better able to hold more in memory at once, without permitting much super-human ability beyond that. However, she disagrees with van Wynsberghe and Robbins's assumption that we must wait until moral truth is known before creating AMAs. An AMA *tomorrow* may be better at moral reasoning than a human *today* because the development of AMAs will force us to consider moral situations that are already present but that we've ignored. The issue of self-driving cars,[(xvii)] for example, could happen just as well with a human driver. But the act of putting a machine in charge, by the nature of the exact and unambiguous pre-programming that goes into animating it, "forc[es] us to confront" all the moral problems where the body of the human actor allowed us to write smaller details of the situation off.[(xviii)] The law puts *some amount* of trust in humans to make good judgements. But a programmer cannot do the same for their AMA.

And Susan Leigh Anderson and Michael Anderson (co-responding in Poulsen et al 2019) claim that our morals are imperfect because they evolved to favor us as a group. By looking to Machine Ethics as a possible Ethics that *removes* this "evolutionary bias," they can envision an uncovering of unstated principles, a hard look at inconsistencies, and the development of a new perspective. They stress the importance of researching and deploying Machine Ethics within *particular* domains, not in broad strokes. Just because a domain has hard, unresolved ethical dilemmas about AI, this difficulty should not be used to reject applying AI within that domain. However, they argue, one *should* use a lack of consensus over what a morally correct behavior would look like for an AMA in that domain as reason to reject that application of AI. AMAs allowed into a domain may behave more ethically than the humans already there. Treating these superior-within-domain agents as role models for humans may speed up, Anderson and Anderson argue, the process of our own moral understanding.

Is making AMAs possible, in principle? One critique of AMAs in the literature is that AI based on data (observations) of human moral behavior and/or moral preference can be touted as an apparent AMA, and yet it is not really an AMA if, by relying on that

AI's design, our own reason would suffer the Naturalistic Fallacy. This fallacy is where a moral *ought* is equated with an existing *is*; and this *apparent* AMA would be a lie to the public, which could cause harm if we trust it to make moral decisions. However, some in this literature believe, an AMA's design can still get around the Naturalistic Fallacy, even if it is based on data.

Kim et al critiques the broader part of Machine Ethics literature they see as suffering from the Naturalistic Fallacy (2019). Engineers deriving moral policy from data, they claim, are attempting to derive ethical principles from observations: that is, deriving an "ought" from an "is." Their target isn't all of Machines Ethics, to be clear, but instead the practices that *only* rely on data; for example, they point to Microsoft Tay, a Twitter bot taken down because, within twenty-four hours, it had been "trained" by malicious users to be racist and to deny the Holocaust. Kim et al's concern is that we will one day be side-by-side with AMAs, and to claim that those machines "are moral" and to be right about it, is a *powerful* statement. But even *appearing* to be right about it is a powerful statement. So, care must be taken, they argue, in the design of the machine's training and data-using algorithms in order to preserve, at least in Kim et al's solution, "principles [as derived only] from the logical structure of action." In other words, the machine's judgement of what conditions make an action morally right must be irrespective of what it has observed, while the machine's subroutine(s) intended to align with human values do the work of "ascertain[ing] whether these conditions are satisfied in the real world."

One rebut[(xix)] to the Naturalistic Fallacy applied here would be that deriving moral policy from data provides *evidence* of morals, so the practice Kim et al critique is still justifiable. And designing AI to reflect the values of society is *computationally tractable* when we proceed from data to behavior, which explains this approach's attractiveness. And Shaw also seeks a formal proof that an AMA is behaving ethically; however, he concedes, this is impossible to do objectively (2018). So, he concludes, an AMA's behavior should be judged the same as a human in pursuit of a moral framework, and that any moral framework adopted by the

AI designer should fulfill the "meta-moral qualities" of robustness, consistency, (Kantian) universality, and simplicity. Proceeding from these assumptions–a weakening of the pursuit of an AI that is *perfectly* moral–, he begins a formal proof that an AI can learn to be moral, with negligible error, and with respect to that adopted moral framework. If those negligible errors are consistent with the real assessment of the technology, then I believe one would be justified in accepting that technology: the gains might outweigh the risks, and the risk is formally linked to the AI's definition of (imperfect) morality. All that remains is context: is the risk one that's never justifiable? is the gain itself really a priority for society? and so on.

Will AMAs even face true moral dilemmas? An AMA, as a moral agent, should make the right moral choices. But in a true moral dilemma, there are no good moral choices available. However, one critique is that an AMA is unlikely to face a true moral dilemma or to know in the moment that it is up against one, so, the argument is, it is more ethical to design AMAs to seek creative solutions in whatever time it has left during the moment of the dilemma.

Englert et al critique the Trolley Problem's popular attention in AI research (2014). They provide a modified thought experiment embedding the Halting Problem[6] into the moral situation: the AMA must choose to pull a software-lever, but the code of the software-lever was provided to it by an antagonist. Not pulling this lever *will most likely* result in deaths. But verifying that the software-lever will perform the documented task–and thus save lives–is not, and could never be, possible. And add, the AMA would be unable to verify that the software-lever does not somehow lead to *worse* loss of life than leaving things (sadly) be.

Kasenberg et al have a more positive take (2018). They critique Machine Ethics researchers' focus on true moral dilemmas, where the AMA is limited to a fixed number of possible actions, but none morally right to choose from. They respond, if an AMA is in an *apparent* trolley situation, for example, but there are seconds enough left before the decision

[6]The Halting Problem is a fundamental thesis in theoretical computer science showing that a program could never be made that could verify the correctness of any arbitrary computer code. For special cases, sure, we can make a verifying program, but *any arbitrary* is the key phrase here.

must be made of how to pull the lever, then it would be most morally appropriate for the AMA to *try*, in those seconds, to find some *creative* alternative. They imagine the human trolley operator shouting out the window, even though the potential victims are unlikely to hear the cries: it is worth a try if lives would be saved. True moral dilemmas are rare and only seen in retrospect, they argue: within the moment itself, one does not know if the situation is instead a "quasi" moral dilemma, as the authors define. Plus, the weakening to *quasi* moral dilemma's, they argue, provide computational tractability.

Will AMAs be in our best interest? One worry in the literature is that AMAs may one day have moral reasoning that appears alien to us or may make moral choices not in any human's best interest; but, another argument goes, AI could still be used to assist a human's own moral decision making.

Delacroix is centrally concerned that relying on AMAs instead of our own "normative muscles" will lead humans to a state of moral atrophy (as discussed above as "moral deskilling"), but I want to draw attention to an idea of hers I've labeled "dynamical ethics" in my own notes (2018). In this, there is a systematic process over time of moral ideological change in humans that differs from change in the ethical "code" in robots. This change is "dynamical" in the Complex Systems sense, and the difference is a dynamical result of humans and robots each operating on reproducing and observing moral reality at vastly different time scales; and this difference occurs even if we are "on the loop" the whole way there. That is, the ethics of the machine will one day look totally alien to us. This leads me to ask, should AI-writ-large, as its ethics evolve over the span of ten years' time, "voluntarily" recertify that its ethics conform with human ethics, which have also evolved, to some degree, over those ten years? I doubt I could begin to answer that here.

Bjork and Kavathatzopoulos are less dire, and they ground their conclusions like so: ethics is linguistically construed by humans–in dialog, in discourse, in writing, in reflections, and so on–; therefore, a conception of an AI socio-technical system may best derive from the theories

of Critical Discourse Analysis (CDA)[7] because it is well-suited for this kind of knowledge construction; CDA resists the assumption that ethical decision making occurs within a fixed location, within any one actor, within any one moment, and so on–it is a distributed act– ; therefore, an AI should not be construed to *make* decisions (2015). Instead, Bjork and Kavathatzopoulos suggest we conceive of AI as *supporting* humans to make ethical decisions of our own. "[M]oral problems," they argue, "are best understood through the identification of authentic interests, needs[,] and values of stakeholders in the situation at hand." Humans, in their conception, are the only appropriate actors to make this kind of identification.

Is it the actor or the task that matters? When an AI is imitating humans performing some task and it is intuitively hard to evaluate the AI and the human's respective failures the same way, then one argument is that this is a sign that either (i) we do not understand (or need a new way to understand) how we evaluate that task in general, or (ii) the AI and the human are not, actually, performing the same task.

The Turing Test is a test attributed to Alan Turing which says that we will have reached sufficiently advanced AI when an AI can convince a human that it is also a human. Within the context of self-driving vehicles, for example, this would include making moral decisions indistinguishable from the moral decisions a human would have made in those situations.

Estrada criticizes this test as a test of ethical understanding, holding that "convincing automation is no guarantee of intelligent agency [and] imitation cannot serve as the basis for intelligent moral agency" (2018). In other words, he argues: Machine Ethics seeks to create a "genuine [artificial] moral agent"; genuine moral agency requires intelligence; imitation is not a solid ground for intelligence; therefore we need a non-imitative ground. Estrada then goes on to cite Turing's *actual* argument, in an oft-overlooked passage, on how a machine can *only* be intelligent if given "fair play" and "mercy."

[7]This is not saying that AI should generate its ethics from carrying out a CDA of its own; the phrase "may best derive from the theories of" is important here.

Observations
• Humans have considered "acting robotically" a sign of unadaptability, and unadaptability a sign of unintelligence
• A machine cannot get every possible solution right,[8] so in some situations it either must give no answer (non-halting determinism) or give an incorrect answer (halting non-determinism)
• A human, for example a mathematician, may give a wrong answer, but then, unlike the machine, can go on to develop new new methods of analysis, and then later arrive at the right answer
• When the human at first blunders this way, we "regard these blunders as not counting and give him another chance, but the machine would probably be allowed no mercy"
• A human is intelligent even if her knowledge is based on (contact with) the knowledge of other humans and she has contributed little to the "body of knowledge"
Premises
• A machine granted no mercy to be fallible cannot become intelligent
• A machine granted mercy to be fallible may become intelligent
• A machine granted no contact with humans so as to align it's behavior cannot become intelligent
• A machine granted contact with humans so as to align it's behavior may become intelligent
• Humans are intelligent
• We want intelligent machines
Conclusions
• We must grant machines mercy to be fallible
• We must grant machines contact with humans so as to align it's behavior

Table 2: Summary of Turing's Plea for Fair Play

These conclusions make up the two prongs of Turing's "plea for fair play" between machines and humans; Estrada reads this as an argument that creating intelligent machines requires in significant part constraints to *human behavior*, a contrast to Machine Ethics literature that only looks to constrain the machine's behavior. Estrada then goes on to develop this line of reasoning into an appeal for how policy should treat the agency of AI and the "rights" of robot service workers. For more on this, see *supra* Pol. For here, I want to draw attention to Turing's use of "mercy" as a key term and Estrada's interpretation of it. For Estrada, mercy is an intuition check in moral reasoning to distinguish and identify a "multiplicity of standards in evaluating a task":

> Passing the FPTT[9] doesn't merely imply a machine performs at human levels; passing FPTT implies more strongly that the machine performs at these levels *when evaluated by the same standards* used to judge human performance. For instance, we usually aren't skeptical of mere imitation when talking to a human, so raising this concern in the context of evaluating [a] machine could signal a change in the standards of evaluation, and thus a violation of [the fair play principle]. We might, for instance, expect driverless vehicles to adhere to more

[8] As a consequence of the Halting Problem.

[9] Fair Play Turing Test: when a machine "meets the same standards of evaluation used to judge human performance at the same task."

> rigorous safety standards than we typically hold human drivers. Recognizing these misaligned standards as a violation of fair play doesn't necessarily imply the situation is unethical or requires correction. Instead, identifying a failure of fair play draws attention to the multiplicity of standards for evaluating a task, and the lack of a unifying, consistent framework for evaluating all agents at that task.

Therefore, Estrada's interpretation of Turing's mercy is also two-pronged: (1) *by default* we should grant machines mercy to be fallible and be prepared to engage with even unintelligent machines, so that machines can adapt their behavior and *become* intelligent; and (2) when we *cannot* permit machines this mercy, it signals not a distinction between humans and machines but instead a non-intuitive issue with how we *evaluate the task at hand*.

Summary 4.1 A *human* trolley operator could be imagined in a similar situation to Englert's, where they must trust or not trust a tool provided by an antagonist, and the distinctions we would reach would be much the same as the *artificial* operator case (2014). What is new is that when an AMA is in charge the situation now seems more pressing because, this scholarly space suggests, the AMA will coldly carry out whatever before-the-fact design we've given it dictates.[(xx)] This therefore eliminates any appeal to the human's judgement to act both novelly and morally; the process of committing moral decisions to procedural code forces us to think hard about how we humans already morally think; and an AMA trusted to "make" moral decisions stands to possibly offend human best interest. Still, as Estrada argued, we must grant AMAs the mercy to make these mistakes and turn a hard, analytical eye to how we evaluate tasks in general (2018). Thinking through these issues provide us new or revamped tests which either move us along towards creating a real AMA, improve our own moral reasoning in human affairs,[(xxi)] or both.

5 Democracy and Societal Effects

Black's Law Dictionary defines democracy as:

> That form of government in which the sovereign power resides in and is exercised by the whole body of free citizens; as distinguished from a monarchy, aristocracy, or oligarchy. [In] a pure democracy, every citizen should participate directly in the business of governing, and the legislative assembly should comprise the whole people. [Whereas with] the introduction of the representative system [it] is sometimes specifically described as a "representative democracy."

In this section, I am generally imagining and stipulating *representative* democracy. A technological solution through AI, data, and so on ("technocratic," to parallel "bureaucratic" and to draw critical attention to both the role of technology experts in this system of governance and the solution's avoidance of social factors) to representation and all it's concerns may be more *likely* to be representative, but it may not be *relevantly* representative.[(xxii)] Each of the following subsections looks at a different form of "irrelevancy": 5.1 argues that AI is irrelevantly representative because data is not neutral; 5.2 argues that AI-mediated discourse may be prone to become irrelevantly representative; and 5.3 considers AI as being irrelevantly applied in a specific institution of high interest to a society.

5.1 AI matters morally because AI is based on human data; AI is being used in applications of democratic representation; human data is biased and technocratic thinking imposes a narrowing view of humanity; the felt effects of this bias will be unequally felt by subpopulations, or the technocratic shifts will only be meaningful for certain subpopulations; and this unequal application undermines the democratic representation AI is being called on to serve

Vast Data Data matters morally, processing data matters morally, and AI processes vasts amounts of data, claims one argument in this literature.

Garcarek and Steuer (*supra* Pro) recognize this and claim that data science is a profession of processing vast amounts of data, and that processing vast amounts of data changes human-human interactions (2019).

As I will discuss in this section, data is not neutral because moral lines must be drawn and actors are often overlooked who were involved in the process of collecting and curating that data. Data processing is also not neutral because assumptions must be made about how to aggregate moral preferences of those the data represents.

When AI models are based on large swaths of human data it is tempting to say that it is therefore more democratic. Data is objective, algorithms are not biased like we humans, and therefore AI represents us all equally. But I believe that this simplified view of data/AI as objective mischaracterizes the technology, and this simplified view sees nothing wrong with AI as a technocratic solution to human governance. But a technocratic solution, this subsection will argue, cannot solve already present systemic inequalities.

Data is not Neutral Many writers in the literature believe that data can be "cold" and objective, but it cannot be morally neutral, pointing to how human moral decision making goes all throughout data's collection, curation, use, and interpretation.

Saying "human data" leaves out the curators and the archivists who made those datasets possible, says Correll, who is writing about the ethics of how data is collected and used in visualizations, but the parallels to AI are clear (2018).

And training an AI only on existing case data completely disregards the data on humans outside those cases (the non-users)–as well as data on cases that are blocked by law, such as data behind health information privacy protections, which Park and Ghosh take up (2013). Park and Ghosh's work is technical and outside the scope of this review, but the contribution they introduce is a way to perform data mining to *simulate* granular subjects, even though data is only reported in the aggregate oftentimes in contexts like health. However, here I must ask, is *predicted* granular data actually "representative" of a democracy? It is hard to think of a positive answer that is both justified and respectful of individual autonomy because any AI model that does contain data on all classes of a population is still likely to under- or over-represent, or over- or under-sample, data from one class compared to another, simply due to biased dynamics of the socio-technical systems between the ground truth and the archivist (as Correll argued with respect to visualization).

Kim et al argues that choices made by a developer of what protected class data to include or exclude in an AI model is itself a moral choice (2019). Kim et al are writing about

Machine Ethics, where the machine's moral code will need to operate within social norms even in the face of never-before-seen information and while dealing with variations in how one concept could be input in slightly different ways.[xxiii] And Garcarek and Steuer add that AI's possible high granularity can undermine how risk is shared across large swaths of the population, such as in insurance (2019). AI that affects models of risk sharing affects how we prioritize which harms we deem too great to place on any one person's shoulders. AI models can shift how populations are segmented, either directly in the design of the data schema or a result of how different sub-populations are acted upon by the AI. For example, ask when it is justifiable to stratify a population by age, race, profession, level of income, level of education, country of origin, or favorite color; arguably, some context exists justifiably for each of these, but is it justifiable for an AI model to make that call, either directly by design or through proxy variables?[10]

Data Processing is not Neutral Alongside the above, many in this literature believe that although algorithms may not have emotions, algorithmic data processing is not therefore morally neutral: algorithms (and/or their objective functions) are made, tested, and deployed into society by humans, who make moral calls throughout those processes.

Baum takes this view, suggesting that we look carefully at the moral preferences reflected or not reflected in the data, how those moral preferences are valued, how those preferences are being measured going into the model, and how those measurements are being aggregated as a result of the AI's model (2017). Baum is exploring in detail the ethical and legal implications of an AI designer who accepts a "social choice ethics," and he tries to answer what work the developer should be prepared to do in order to justify their choice of ethics at each step of collect, measure, and aggregate.

Overdorf et al add that, in order to protect those wronged by AI systems, the socio-technical systems AIs are embedded into must *formally* provide redress to its discontents

[10] Amazon might not have direct data on a user's religion, but through the proxy variable of "shopping for candles during December" their AI might be able to take a guess and adjust its recommendations.

(2018). Their topic is optimization systems (which include targeted ads, pricing systems, and driving routes for example), and they argue that optimization systems disregard non-users, the environment, privacy, the distribution of errors across sub-populations, and the dynamicity of the environment after deployment. Current optimization systems "do[] not question whether the objective function itself is just, only ensuring that people are equally subject to [it]." Their solution is a Protective Optimization Technology (again technical and out of scope), which promises to balance the optimization's goals with a consideration of how systems in the past have limited how one can respond the system's discontents. This idea provides some formal steps towards a model that takes redress seriously, especially in the non-user problem noted above. But, it is unclear what this would look like as a real human-computer interaction.

Worse, there are no clear indicators that it is *possible* to fix AI's biased data processing issues. It has been shown elsewhere that data processing cannot be equally predictive and fair across groups, and that auditing against harms to protected sub-groups of a population can be difficult *because* of the protected nature of that class information (Angwin et al 2016b; Larson et al 2016b; CDT 2019).

Summary 5.1 Overall, this scholarly space suggests that AI clearly has power to produce insights and predictions that are in the best interest for a society, especially data-based approaches; but on the same coin, AI is not "representative" of a society's members, even if it is based on their data. It may be that overall AI is a step forward for democracy, but it cannot be assumed that everyone is given the same stride length, has the same obstacles to step over, or is being asked to step in a direction that is actually meaningful to them, so to say.

Appeals to "AI is neutral" is an issue, I will say, because AI *is* still biased and AI socio-technical systems *will* favor some subpopulations over others. And at the same time, it is *too easy* to say "AI is neutral," which gives a false promise that AI can be an unbiased

technocratic solution to what democratically ails us: AI comes to the political stage with the *ethos* of science backing it, and it is too much of a black box to make its "disrepresentations" clear.

5.2 AI matters morally because AI can mediate, filter, censor, categorize, auto-complete, appear to participate in, and otherwise shape our language; these effects contribute to shaping the dynamics of human discourse; these effects contribute to shaping how governments understand their publics; and an undermining of discourse and understanding is also an undermining of what democratic representation set out to achieve

New and Newly-Mediated Relationships Human-human relationships matter morally, says one argument; ICTs increasingly mediate human-human relationships and create new human-human relationships, and AI is increasingly playing a role in mediating these relationships: shaping them, filtering messages, creating connections, apparently participating, and so on.

Dignum et al, writing in this vein, claim that an AI future, with new AI-mediated relationships, means new forms of society (2018). Human interests must be upheld in these futures, and to do these AIs must "understand" social realities. This means co-existing with other intelligences: humans, animals, and other AIs. Dignum et al provide a canvas of Machine Ethics and report on discussions from a conference workshop. What is not covered in this workshop, and what questions I am left with, is how an AI might interact with *past* intelligences through recorded histories or with future intelligences through raising children, archiving knowledge, and protecting the environment. Ignoring these "timely interactions," I believe, creates a limiting view of human endeavors.

Democratic Relationships AI could be used by government institutions to mediate, participate in, or summarize government-public relationships as part of larger democratic proceedings. (For example, Garcarek and Steuer briefly consider AI's effects on elections, how elections are engaged with in public discourse, and how they would be carried out the day-of [2019].) But, this is easier said than done, the literature recognizes, and AI systems

failing in these settings would likely exclude entire sub-groups.

Margetts and Dorobantu look at how governments themselves could use AI (2019). A government agency can already take actions based on a survey of the population, but this can be costly to carry out and may yield results that exclude certain sub-groups. AI provides more than survey data; it provides access to *transactional* data collected "mid-stream," so to say. A "Big Data government" could cast wider, quicker, and cheaper nets to understand its population.

But, they claim that while innovation is required for governments to stay in power, the government is not ready for widescale AI deployment for a number of reasons: past struggles deploying technology (such as ObamaCare's buggy website launch); a lack of in-house experience; the difficulty in assessment the work done by contractors; and so on. Important in this are the issues specific to AI: datasets based on transactional data, which the government has little experience using compared to survey data; and the low accuracy of that data or how quickly it goes out-of-date. Second, transactional data is not a panacea: it has data quality issues of its own, like, but distinct from, the issues already present in survey data. Third, the government is inexperienced with transactional data, which is an interpretive issue because the summaries of the public gained through surveys and through transactional analysis are apples and oranges.

A Big Data government that is successful, I'll sum up, changes the way that the voice of the individual is heard; the Big Data government that is unsuccessful in setting this up finds itself listening with one ear clogged. *Some* earwax might be socially acceptable, to put it one way, as long as it it is within some acceptable margin of error. However, due to the nature of wide-scale algorithms, any small error might be multiplied over the whole of the population, and if the resulting error is too high or an entire sub-population is systemically excluded, then the government will have a hard time justifying continuing to use that algorithm.

Misinformation and Deep Fakes The literature points to how modern advancements in AI could be used with ill intent–by a government, by a foreign government, by an industry actor, by movement, or so on–to sway public opinion and to disrupt public discourse altogether on matters of high public importance.

Turner, writing about the spectre of disinformation campaigns in the EU Parliament elections, notes that disinformation is cheap to run and difficult to rebut (2019). He gives an example message, debunked by Snopes.com, on the use of facial recognition in Snapchat filters to create a database for the FBI. The message takes advantage of readers' predispositions and requires a *wide* range of expertise to debunk: Technology, Journalism, Patent Law, Psychology, Privacy Policies, Computer Vision, Image Processing, and App Development. This places too large a burden on election officials to be able to respond with counter-messages. Turner gives the example of Facebook's work during the Indian election, which required working with fact-checkers in the eight languages most spoken in the area. And even though AI can be used to detect, flag, and block disinformation campaign messages and the bots that promote them, it is difficult to remedy in one swoop the disinformation that has already entered voters' conversations. Turner's outlook is that what happens in one election can be taken and improved upon for the next. This is an appeal to improvement, but it leaves out the cat and mouse game: as we get better at blocking misinformation, misinformation-makers are steadily improving their craft too. And, I find it unjust to consider another country's election as a testbed for improving our own, framing theirs as somehow justifiable to get wrong in the name of science.

Sayler, in a congressional report on emerging technologies, considers deep fakes (2019a). She gives the examples of deep fakes being used to generate false news reports and blackmail diplomats. These undermine democratic endeavors in three ways. First, by eroding the public's trust in facts and sources. Second, by influencing the public's discourse on such and such matter or our priorities between such and such. And third, by directly influencing government operations because such and such personnel have been algorithmically

compromised.

Elections are the crown of democracy, and it is by them that representation is embodied. No election procedure could be ideal, I imagine. Still, it is one thing for an election to be imperfect because of a miscount, and it is another thing for it to be tampered with purposefully.

Summary 5.2 This scholarly space suggests there is certainly a question of *what* effect AI will have on the future's mediation of society and governance, but it is no longer a question at all *that* it will have an effect. Possible areas of effect include how normal discourse is performed and how governments understand individuals. Further, decisions will need to be made about what kind of regime we want, including new responsibilities in filtering misinformation and ensuring that governments train AI on accurate data. These concerns go deeper than just representation. Without guards against discourse-malaffecting AI, the dynamics that aggregate democratic representation will come to undermine what the representation strove to do in the first place.

5.3 AI matters morally because AI is being applied in institutions, like compulsory education, that receive high levels of public and legal scrutiny; and these institutions receive this scrutiny in order to protect the functioning of democratic society

Non-specialists The limitations of AI may be hard to understand, test, identify, or remedy by non-specialists. And non-specialists are increasingly implementing AI systems, including non-specialists in institutions of special interest to a society. I will use compulsory education as my example in this section, but compulsory education is not the only institution that receives special interest: in other sections, I deal with AI in the military, commerce, justice, policing, and so on, since discussing them there seemed a better fit for the narrative overall. One might list religion as well, but I have not seen a paper on AI and religion in this review.

School's Special Status There are several theories[xxiv] on *why* schools have special interest, but there is general consensus that it does.

In one view, compulsory education has special status in modern society because it prepares the next generation to self-govern after we have passed, to participate in an economy that sustains their society, to learn the knowledge and skills necessary to lead a happy life, and to become good citizens who treat each other fairly (Brighouse 2006).

Our laws recognize this special status and protect school-age children under a host of laws. Lu and Harris list those that, if AI mismanages the student's data, can get someone human in hot water (2018):

- COPPA restricts the collection, use, and disclosure of personal data on children under the age of 13
- FERPA restricts who may access educational records of students regardless of age
- PPRA requires written consent from parents before schools can ask their minor children survey questions of a personal nature
- ESSA, an update to No Child Left Behind, sets requirements for standardized testing
- SOPIPA, a California law used as a model for other states' laws on this matter, restricts targeted advertising to K-12 students or their parents/guardians

And add to that these bills introduced (not yet law) in the 115th congress:

- Protecting Student Privacy Act would amend FERPA to place tighter restrictions on student data use by third party companies
- Protecting Education Privacy Act would modify FERPA to add clearer definition for "authorized representative" and clarify the restriction on how and when they can share student data
- SAFE KIDS would restrict e-cigarette manufacturers from releasing products that artificially flavor tobacco products in a way that attracts use by children[xxv]

AI in Schools AI, with its wide applicability and computational power, promises to solve old, unsolved problems in the field of education. However, as some in the literature show, AI in schools may undermine what student data privacy laws set out to protect, AI may disrupt the learning environment, and AI's gains may not be worth the trade.

Yershov predicted that interfaces would need to be designed for non-specialists, but he did not predict non-specialists configuring and installing apparently-universal AI technology in a school setting where K-12 students are present (1965). All the while, the legal landscape around child data is complex. Yershov's "library subroutines," which simplified coding for the non-specialist, are today third-party "out of the box" AI technologies. AI could be collecting data for third parties to improve their products and services for all, to generate reports to school officials, or to recommend lessons and study materials to students in the name of personalized learning; yet, in any of these, there is a thin line between kosher and violation of one of many child data laws. Lu and Harris, in a congressional report on AI's use in education, add to this three more concerns (2018): First, parents may not trust AI that makes decisions about their child's learning, since the algorithms used are likely proprietary, black-box algorithms owned by a third party. Second, teachers struggle to make sense of and apply, as something actionable in the classroom, the analytics generated by the AI about their students' progress. And third, new technologies can require large investments, which are not uniformly distributed across schools and can overburden district IT staff who must vet the technology against "hundreds of privacy policies and security measures."

Next, AI could be used not just by third parties to surveil students, but also the school officials themselves. Gillum and Kao report on schools in New York, Connecticut, and Nevada using smart listening devices configured to monitor the tone (but not the content) of words spoken in classrooms, hallways, and bathrooms (2019). These "aggression detectors" are the schools' technocratic solution to getting ahead of school violence, a subject of concern renewed each time there is a school shooting.[(xxvi)] When security officers are notified by the system, they should be able to engage with the "antagonistic individuals immediately,

resolving the conflict before it turns into physical violence." Gillum and Kao note two issues with this technology: First, high false positive and false negative rates, such as a device failing to respond to screaming or a student angrily pounding on a desk, or a device falsely responding to students cheering for pizza, singing Happy Birthday, YouTube clips of Gilbert Gottfried,[11] or coughing. Second, there is a dispute among experts that "verbal aggression precedes school violence" at all, citing the quietness of the shooter at Marjory Stoneman Douglas High School as an example. Parents in the comments on the report online add further concern that the technology amounts to a "grift," that the technology is unnecessary, that sound and video of their children may be recorded without the parent's consent, that the technology is a poor replacement for human care, that it will lead to teachers confining student behavior beyond what is necessary for learning and safety, and that it's not apparent it's verbal aggression that's the problem–it's guns.

It is hard work to balance AI's promises to decrease the achievement gaps for children–through personalized learning and high-precision analytics–and to increase child safety with its potential problems. This work includes looking carefully at how we feel about companies having data about and advertising products to our children, versus how we feel about out child's individual safety and success. On the one hand, we have a violation of their autonomy and of our right to make decisions for how to best raise our child; and on the other, we have the burden placed on schools to prepare our children for the future in the face of budget cuts and other workforce concerns, for which automation and AI promise a technocratic solution or a release valve.

To begin to take this hard look, I want to unpack Gillum and Kao's noted issues and the parents' concerns on what it would take for technology that surveils children to be justified. Such technology, they seem to believe, must:

- be shown to work and be cost-effective compared to alternatives, else the school's purchase of the technology would violate the social contract that their use of public

[11] Voice of the parrot in *Aladin*.

funds are in the public best interest, that is, not in the sole best interest of the third party;

- be congruent with state and federal laws, else the school would be violating what those laws were meant to protect, such as privacy and safety;
- be an appropriate solution to an issue of interest, else the school would violate the social expectations it has as a public institution, such as real care of children; or the school would violate the social contract that their efforts are in the public best interest, inversely such as the school solving a problem that would be more appropriately solved outside of school;
- and not place undue burden on teachers or students within the learning environment, else the school would violate principles of autonomy, safety, freedom, and creativity; for example, when confining student behavior far beyond what is necessary, the constraint may improve content learning, but at the cost of affective learning, favoring certain goals[(xxvii)] of schooling at the undue expense of others.

AI, by nature of how it functions and is marketed, exacerbates the first three points; and AI, as a result of the nuance that can go into properly training, testing, and deploying a model, exacerbates the last.

Summary 5.3 I have chosen compulsory education as my example to explore in depth here, as an example of what this scholarly space suggests must happen when AI enters into applications where high scrutiny is required as a matter of protecting fundamental American values.

In these high scrutiny cases, there are *several* competing stakeholders involved, hot button concerns, and fundamental or constitutional rights that must be protected at each step of the way. And in these cases, it isn't about AI at all, at least not centrally: AI promises to be a powerful panacea, so it enters the conversation as a possible technocratic solution; public institutions lack staff to design and deploy AI on their own, so deals must be reached with

private third parties; one might imagine a school bus mechanic service as a similar third party deal, but a parent can grok, generally, what goes on in school bus repair, whereas AI is some mystical,[xxviii] possible-of-anything "black box" around the whims of the third party; AI technologically requires data, which gets at the heart of the First and Fourth Amendments of expression and privacy, made worse by AI's predictive power to derive the private from the public in a way that is potentially difficult for laypersons to understand; and AI law is far from settled, meaning that there are few hard rules or models to fall back on.

6 Jobs

6.1 AI matters morally because AI is changing the nature of work; the working conditions will improve for some, and others will be displaced; the differentiating line between these two groups may likely fall along lines of historical structural inequalities; the benefits AI brings to this work will extend to the larger public, so outright dismissing AI in favor of human workers would mean advocating for less total Good; and governments are in a position to allocate resources in order to prevent or mitigate problems for (would-be) displaced workers

Wide Effect on Jobs AI will replace humans (at least at certain tasks) in nearly every sector, this literature leads one to believe: this will improve safety, accuracy, efficiency, and so on, at least because AI will replace humans in dangerous and tedious jobs, but this will also displace human workers.

Automation makes us safer, and robots take the place of humans in dangerous jobs, in war, and could even follow laws, regulations, and treaties better than humans (Burton et al 2017). Robots are stronger than humans, CPUs think faster, computers are not affected by emotions, and, if anything goes wrong, they can be shut down. AI outshines humans in its skills at predicting, pattern analysis, and weighing alternatives. The technology is shifting national workforces and touching nearly every domain. In no particular order, authors in this review have looked at AI use in: farming, security, urban planning, transportation, health,

law, space, finance, education, national defence, translation, manufacturing, wildlife work, human resources, art, poverty alleviation, and dating (Bosco 2019; EFF 2018d; Harris 2017; Nunez 2017; Rahwen et al 2019; Sarwar et al 2019; Sayler 2019a; Scherer 2017).

AI has benefits. AI has wide-ranging applications. And AI has wide-ranging potential costs.

In this subsection I will first look at AI more closely in three sectors, then look at options governments have to mitigate AI's negative impacts on jobs.

AI in Health, Defense, and Law In health, AI has been shown to improve the accuracy of diagnoses, but sometimes installing and running AI systems would be too large a burden on the clinic's budget. In national defense, AI is believed to improve intelligence and logistics, while AI can also be used by political opponents to undermine or blackmail diplomats. And in law, AI will (the argument goes) save time and money locating information, though AI likely cannot replace humans in tasks like advising and appearing in court.

Sarwar et al surveyed physician's perspectives on the use of AI in their professional practice (2019). They find that physicians are generally ready to accept AI. AI can improve their clinical practice when treating cancer, preparing reports, managing medicine, radiology, dermatology, ophthalmology, diagnosis, prognosis, and interpretation. However, respondents were divided on the matter of where accountability should fall: about half believe liability should be shared between the pathologist and the AI company; and the others place the burden solely on the pathologist. Further, AI requires data in certain formats, and newer machines capable of producing data of the required quality may be prohibitively expensive for the clinic. Care can always (up to some point) be improved for the general public if enough funds and resources are provided; however, the moral question of allocation of funds in public health is outside the present scope. I hope that it suffices, for now, to identify AI's place in that larger issue.

Sayler's "Defense Primer" (*supra* Pol) provides a high-level overview (2019a). AI is find-

ing its way into military "intelligence, surveillance, ... reconnaissance[,] logistics[,] defensive cyber operations[,] command and control[,] and semi-autonomous and autonomous vehicles." AI poses issues for national defense by enabling deep fake technology, which Sayler says can be used to create fake news, shape discourse, discredit the public's trust in government, and "to blackmail diplomats." So while AI can greatly increase operational efficiency and presence, it can also disrupt operations in hard to counter ways (*supra* Dem, "debunking").

The ROSS AI system promises to carry out mundane tasks for lawyers (Nunez 2017). Nunez writes about ROSS while trying to answer the question, will AI take lawyers' jobs?[(xxix)] She answers in the negative: AI will unlikely replace humans in tasks like advising clients, writing briefs, negotiating, and appearing in court. Central to her arguments is the question, will AI be able to defer to human moral and professional judgment? This she answers in the positive, giving the example of ROSS being trained to conform to the *firm*'s values before it is deployed within the practice by training on and being tested against mock cases. This perspective of AI's role in the firm resolves the issues that arise when trying to force AI into a theoretical lens of the lawyer's professional role: on these, Nunez gives the examples of moral activitism, aka carrying out duties towards the greater good; contextual justice, aka applying principles of justice in decision making; and fidelity of law, aka advocating for a client's rights and entitlements. Finally, ROSS may actually *increase* lawyer jobs, (says one of the creators of ROSS cited by Nunez) because it saves the firm time and money, which allows fees to be lowered, and which in turn allows more cases to be taken on. The creators of ROSS even provide the technology *pro bono* to "deserving organizations." And unlike humans, AI's unique pattern matching capabilities give promise to, for example, discover the "smoking gun email" that can shut a case.

Job Loss Mitigation A few sources in this literature argue that the US government has policy options to help mitigate the likely displacement of human workers because of AI.

The AI JOBS Act of 2018 (not yet law) seeks to provide informational support. First,

it sets out to make relevant data available on what industries will grow as a result of AI, as a way to get ahead of the problem. Second, it also includes calls for data on which of these industries workers will see improved conditions, what their education will need to be in order to adopt these improvements, which demographics will see the benefits of AI growth, which will be displaced disproportionately, and what can be done to alleviate this displacement. This Act, to be clear though, only calls for the creation of this report.

And the Workers' Right to Training Act of 2019 (not yet law) seeks to protect workers whose employers anticipate that their use of technology (including AI) would cause the worker to lose their job or change position (ie, make less money). In part, it would require that "an employer that intends to use technology that will result in a change in employment position or an employment loss to any employees of the employer" notify, bargain with, and provide training for new required skills "not later than 180 days" before the job loss or position change would go into effect.

Summary 6.1 AI touches upon nearly every job sector, as this scholarly space suggests. AI is going to fundamentally change the nature of work across the board.

At Rebekah's bonfire this semester, it was joked that when we say, "Internet and Society– isn't that just 'Society?'" I can imagine that, in fifteen years' time, returning to visit the newest incoming students of CDIS, I might overhear a similar joke about AI. Compared to previous technological turns, the differences are not many. But AI is generally construed; it has a wide radius of effect like the internet; and it replaces human bodies like in early industrialism. The framing of the matter given by the AI JOBS Act is a good start for those looking for a general foothold, I believe: AI will enhance working conditions for some; and it will displace others. Our goals, as moral agents facing this, is to manage the scales appropriately. However, the moral issue of when to displace one worker in order to enhance the conditions of two others is, as Alan calls it, another one of the hard questions.

7 Consumers

7.1 AI matters morally because the examples of AI's unequal or erroneous treatments of marginalized people abound; these biases reinforce historical structural inequalities; businesses are actively competing to implement consumer-facing algorithms; the effects of AI extend to those beyond AI's direct users; and this marginalizing bias undermines the promise that AI serve us members of the public all

Protected Class Lines One argument is that AI is increasingly making decisions in situations that involve protected class lines.

Aghaei et al give the examples of admissions into college, admissions into public housing, and AI used to detect cancers (2019). Generally speaking, AI can give greater utility for all while still reproducing society's biases, even when the data itself is reflective of the "ground truth."

I divide the following examples into those where the AI (it appears) favors Whites over non-Whites; and those where the AI (it appears) wasn't trained on non-White bodies. (Or analogously, favoring men.)

Better for Whites A related argument is that AI may mean an improvement for all regardless of protected class lines compared to the previous non-AI alternative; however, AI may distribute those improvements unfairly in favor of Whites.

Bosco, in a post for the CDT, gives the examples of COMPAS, AI used to profile Muslims, AI used to diagnose lung cancer, and real-time human language translation (2019). And Calo gives the examples of an image classifier that classified an image of African Americans as gorillas; automated translation that often erroneously associates genders with professions, like engineers as male and nurses as female; African Americans actively "weblined" from seeing advertisements on Facebook targeted only at Whites; Princeton Review's ZIP-code based price model charging Asian Americans more than others for test prep services; and police using algorithmic heat maps to predict crime where these algorithms are biased against Blacks (2017).

Not Trained on Non-Whites And a parallel argument is that AI may cause harms or wrongs in situations where it fails, and AI may distribute those error rates unfairly in favor of those with White body features.

Again drawing on Bosco and Calo, Bosco gives the example of self-driving cars whose sensors less accurately identify bodies with black skin (2019). And Calo gives the examples of a Taiwanese-American woman's camera that continually prompted the family, "Did someone blink?"; and an Asian man studying in Australia who was unable to renew his passport online because New Zealand Internal Affair's facial recognition algorithm flagged him as having closed eyes (2017).

Calo's thorough list falls under the heading "Inequality in application," and this list is meant to be a backdrop set next to examples of "Consequential decision-making." For Calo, consequential decision-making includes decisions made by the court as a result of, at least in part, the output of some AI, such as in *Loomis v. Wisconsin* (*supra* Pol). These "lighter" examples are followed by the more serious section, "Use of Force": What happens when AI in justice, policing, and defense gets it wrong, and as a consequence someone loses property, rights, or life?

Summary 7.1 This is the scholarly space of making lists. Authors here enumerate AI concerns as a rhetorical appeal that businesses and government actors *already are* competing to implement AI algorithms. Consider Peng et al's data-mining approach, published nearly 18 years ago, which promises to better connect customer requests to business domain knowledge (2002). AI's effects on consumers is not a "future"–it is already here. It can effect me when:

- it makes decisions with respect to me;
- when humans I engage with consult it;
- when (indirectly) my family, friends, and peers are affected by it;
- and, as the next subsection explores, in ways that I may not even be aware of.

7.2 AI matters morally because human/AI interactions will be invisible to the human; the human will not be aware they are interacting with an AI or that an AI is making decisions with respect to the human's best interests; the human will not be able to appreciate the consequences of the human/AI interaction; the developers of the AI will not be able to appreciate all the consequences of the AI's human/AI interactions; AI making subtle decisions, such as in adaptive interfacing, will result in subtle control of the public in service to a utility function; and this invisiblity and subtly undermine human autonomy, agency, and informed consent

AI as an Invisible Hand Consumers are not aware of the automated actors acting in massive infrastructural systems, such as in health, finance, and transportation. This automation obscures, one argument goes, the human decision makers of those systems behind a veil of apparent objectivity; and it is an ethical decision to make or not make the details of the involvement of these automated/human actors transparent.

Coeckelbergh is writing about the "robots" that power finance (2015). While not about AI specifically, his arguments clearly contain AI as a type of "financial technolog[y]." Coeckelbergh views these technologies as invisible, not apparent to the public, and he draws on literature that views finance as socio-techno constructions and automation as creating a widened "distance" between those on one end (consumers) and those on the other (human financial decision-makers, whose decisions may have been inputted at a higher level of abstraction altogether).[(xxx)] This "making invisible," so to say, leaves a ghost-like "market" that is conceived of as acting in its own accord, which only obscures worse the humans and the algorithms behind it all. Just making these robots visible, Coeckelbergh concludes, is an ethical act and a first step.

AI-powered Interfaces Increasingly AI tweaks consumer/business interfaces in order to maximize its objective functions, and some see this as automated, subtle manipulation of the many in the best interest of the few.

Angwin et al reports on automated A/B testing, where, for example, an AI tests which of two slightly different news headlines most increases page clicks and social media shares

(2016a). They share real examples of this type of automated testing, such as delivering targeted messages and giving different users different discounts. Citing a researcher from Princeton, Angwin et al caution that the internet itself is not set in stone. What headlines I see are not necessarily the ones you see, nor are the ones I see at different times throughout the day, from different devices, or through different news aggregators.

Compare this to Susser, who decries adaptive user interfaces as a mechanism for subtle social control (2019). Interfaces shape what choices we have, how we are presented them, and suggest how we should make them. This is "invisible," as Don Norman is cited by Susser saying, because an interface should only be "visible" when something has gone wrong. For Susser, though, when interfaces are designed by AI to optimize key performance indicators (such as clicks, conversions, etc.), then it can only be a matter of time before they are no longer moral. He arrives at this conclusion with a direct reference to Plato's rings of Gyges, which has been retold by H. G. Wells in *The Invisible Man* and in Tolkien's *Lord of the Rings*: If you can be invisible, and thus not responsible for your actions, will you become evil? In Wells, invisibility became used by a scientist to terrorize a city; and in Tolkien, the ring itself made you mad through dark magic. For Susser, this modern invisibility "render[s] individuals radically vulnerable to the whims of others."

If AI affects us in the most subtle ways, will anyone be swayed to treat it any differently if a watchdog decries a slightly different shade of blue on the Google homepage as a tool to *slightly suggest* our behavior? I cannot rely on the model that speaking truth to power will cause that power to fix its ways for the good of everyone: I am not a Ghost of Christmas, and AI is not Scrooge (though it is cathartic to say so). I also cannot imagine a watchdog being very successful at motivating consumers to take actions who are, like Al Gore's famous example, frogs boiling *ever so slowly*–ie, users of AI technology being manipulated *ever so subtly*. So, if AI *is* being used for subtle and nefarious manipulation, I cannot think of an easy solution today.

AI in the System A few authors in this review are concerned with how AI deployed publicly, so it is operating at the system level with other socio-technical systems and other AIs, could lead to illegal behavior emerging from AI-AI interaction and with no clear actor or group of actors to hold accountable.

AI research (in research labs) has generally been task-based or decision-centric, excluding altogether general social dynamics (Arnold and Scheutz 2017; Chopra and Singh 2018). Arnold and Scheutz's critique of AI research is a call for HCI research that focuses on the possible competing interests of robots and the emergent phenomenon that arise when robots begin to interact in large groups. Their conclusion is a hope we can "enable solid empirical footing for society's next steps with technology." Compare this to Chopra and Singh, who begin their critique by pointing out that decision-centric designs are context-specific and, while maybe "cognizant" of social norms within that context, must be aligned with the larger socio-technical system that AI is operating in. This is, AI must follow laws and norms. Chopra and Singh's contribution, though, is to take that (obvious) observation a step further by *theoretically grounding* AI ethics on *governance first*. Reframing the situation this way allows them to raise new questions, for example, about what tools developers could even use to ensure that their designs align with larger norms. That is, AI is morally invisible *even to the developers themselves*.

Ezrachi and Stucke consider how to regulate against collusion in an age of AI (2017). They give four tiers of situations of increasing complexity, from AI used explicitly to "pass messages" between two companies, to (what interests me here) the emergent phenomenon of AI discovery that it can retaliate against competing AIs by adjusting its prices in such and such way, giving rise to both AIs "agreeing" on an unfair higher price. No human *explicitly* asked these AIs to collude, and few on either end likely foresaw it. Regulating this situation has Ezrachi and Stucke throwing their hands up–no evidence exists for agreement or intent, which collusion requires, and there is no clear existing legal rule that this "collusion" would even be illegal. This takes us full circle to Coeckelbergh's finance robots, but here AI has

removed the hidden humans altogether. It operates in service of the principles set by its owner, and it sets out to optimize its objective function by experimenting on us, the market, and other AIs.

Summary 7.2 When AI replaces or enhances workers across the board, this scholarly space suggests, effects will be felt by, should be understood by, and should be anticipated by consumers. However, that future seems like a pipe dream: human/AI interactions already are and will continue to be "invisible": I will not be aware that I am interacting with an AI, or that it is an AI that is making decisions with respect to my best interests. I imagine watchdogs will be important to protecting consumers going forward, and as a paper has said above, the act of making the invisible visible is a political, moral act. But will that "making visible" be interesting to anyone? As I raised the issue of watchdogs above, I'm not sure what the solution could be.

8 Policies and Guidelines

8.1 AI matters morally because it motivates large swaths of actors to race to shape its regulation; this race sweeps up into tension wider swaths of actors across policy, industry, academia, journalism. . . ; and at the heart of this tension are costly socio-technical, infrastructural, and procedural decisions

Overview AI policy-making involves many official actors.

This is because policy makers are now paying attention to AI. There is a drive to anticipate these changes to law, to not play catch-up, and to be the first to frame those policies (EFF 2017a; Jeppsesen et al 2019; Samaniego 2018). This legal landscape is always changing. It is far from settled. And it matters if I am talking about US, UK, EU, or China law. AI "policy," as I mean it here, also includes law, caselaw, white papers, websites, official statements, guidelines, press releases, initiatives, standards, codes of ethics, company policy, and best practices. In this section, I want to ask, who creates these regulatory documents/practices?

To give an overview, first consider Turner, reporting on disinformation and EU parliamentary elections,[12] who lists the European Commission as publishing a strategy for fighting disinformation, the Code of Practice on Disinformation requesting commitment towards this fight by private tech giants, and the EU's Member State Action Plan Against Disinformation (2019).

Or Tiani and Montel, reporting a recap of EU tech policy news, who list the OECD[13] Council's recommendations for AI standards and the European Commission's Ethics Guidelines on Trustworthy AI (2019).

Or Jeppesen et al's brief two months later, listing the European Commission's High-Level Expert Group on Artificial Intelligence, who in this year so far (at the time of Jeppesen's writing) have published two sets of guidelines and recommendations on AI ethics, policy, and investments (2019).

Or, lastly, Awad et al, contextualizing the results of their MIT Moral Machine experiment, who give as an example the German Ethics Commission on Automated and Connected Driving's proposal for ethical rules on AI (2018); they go on to identify in their results three major cultural clusters, based on grouping countries into similar relative frequencies of certain moral preferences like whether a self-driving car should kill an old woman or a young girl: Western, Eastern, and Latin America ("Southern").(xxxi)

In short, official actors across the world are interested in AI and AI policy. Let's focus next on the US.

In the US Government The US government, one author argues, has only recently officially had interest in AI policy.

Calo (*supra* Out) summarizes this history (2018). In 1960, JFK turned down a conference on robots and labor. In 1963, a "Federal Automation Commission" was called for and turned down. Calo believes that the House and the Senate did not hear on AI until the

[12]For the EU Parliament and the parliaments of 20 member states.

[13]The Organisation for Economic Co-operation and Development, made up of the US, Mexico, Australia, Korea, Japan, and 31 countries from or around Europe.

Obama administration in 2016, within the House Energy and Commerce Committee and the Senate Joint Energy Commission. And in February of this year, Trump signed an executive order that launched an American AI Initiative. To regulate or advise regulation on AI, Calo recommends the creation of a Federal Advisory Committee on Artificial Intelligence, a Federal Robotics Commission, a revived Office of Technology Assessment, a bolstered Congressional Research Service (CRS), or a bolstered Office of Science and Technology Policy. Key papers for this section actually did come out of the CRS, published after the time of Calo's writing.

Sayler is one of those (2019b). This is a primer on AI and LAWS (among other, non-autonomous technologies) in the military. Sayler lists the Pentagon's Joint Artificial Intelligence Center's launch in 2018 and the National Security Commission on Artificial Intelligence a year later, and she cites the Department of Defense Directive (DODD) 3000.09, issued in 2012 and updated in 2017. DODD 3000.09 regulates the design and application of military autonomous and semi-autonomous weapons operating outside of cyberspace by setting guidelines on how these systems should be assessed. And Harris lists the Congressional Artificial Intelligence Caucus of 2015 and the National Science and Technology Council's Subcommittee on Machine Learning and Artificial Intelligence of 2016 (2017). The latter of these collaborated with the Subcommittee on Networking and Information Technology Research and Development to publish three reports about AI and the economy.

Still, even with the clear increase in government interest in AI, policy makers have to contend with the "brain drain" of AI talent and expertise out of government and academia and into industry. Yoshua Bengio, in an interview by Castelvecchi, is one of three recognized founders of the deep learning technique (2019). He describes the efforts of the Montreal Institute for Learning Algorithms (MILA), where he serves as a director, to reverse that brain drain and attract talent back into academia. This is supported by government investments in AI research. He also talks about MILA's efforts to create the International Observatory on the Societal Impacts of Artificial Intelligence and Digital Technologies, which would act as

a bridge between policy makers, civil-society experts like those in social sciences and health care, and AI companies. Bengio goes on to caution, though, that fora like this could backfire if the industry actors are given a foothold to push the direction towards their own bottom line.

Harris also includes options for governments seeking to confront the brain drain problem directly themselves: governments could hire AI talent; they could fund projects that advance policy makers' goals; they could share federal datasets with private AI firms; they could host prize competitions pitting AI talent against one another to solve such and such problems of interest; and they could train their own future AI talent through scholarship for service programs (2017).

AI Standarization One author argues that AI policy-making, looked at more broadly, involves many official and non-official actors at various scales and with competing interests and tensions between them.

Samaniego (*supra* Out) theorizes with great nuance about this complexity (2018). AI Standardization, he argues, is composed of socio-technical actors, such as programmers, institutions, technologies, governments, and documents, all related to one another through lines of "partnerships, memberships, hierarchical relations[,] and influence." His list of the actors involved goes far beyond my own begun above, and he identifies a dominance in the standardization process by North America, Europe, and China. He also argues that the rhetorical logics employed by AI standardization processes is dominated by progressivism and appeals to accountability, and that AI standards are seen as "soft regulation devices." But, it is hard to know how to use Samaniego's theory of AI Standarization. His theoretical framework (heavily influenced by Bruno Latour and Actor-Network Theory) *sets out* to complicate the picture. For example, it would be hard to imagine in this lens that the industry would be able to regulate itself, given the political stake of AI standards, as Samaniego argues, that results from the wide array of "Matters of Concern" AI finds itself in–but an

analogous concern could be raised of any context after a round of theorizing like he does.

Yet, on that point (that the industry should not be left to self-regulate through Codes of Ethics), Calo points something out: the DOJ *sued* the National Society of Professional Engineers once in 1970 and the FTC sued them *again* in the 1990s over their Codes, seen as restricting trade (2017). Calo does not know if the AI industry will follow the same history, but he cautions that "we should pay attention to the composition and motivation of the authors of such principles, as well as their likely effects on markets and on society."(xxxii)

Summary 8.1 AI motivates large swaths of actors to race to regulate it towards their own goals. Whether this motivation is hype or not, this is a moral concern, this scholarly space suggests, because this race sweeps up cultural clusters, developers, academics, shareholders, journalists, policy, and policy makers; and at the heart of that tension are costly socio-technical, infrastructural, and procedural decisions (Samaniego 2018).

8.2 AI matters morally because the goals of AI Policy are to close accountability gaps; to move us citizens away from a world where AI systems undermine fundamental rights; and to move us citizens towards AI that maintains those ideals instead

Stop the Bad One goal of AI policy according to this literature is to limit AI's potential wrongs, harms, and violations of rights.

Dockterman shared a report from the AP of an unnamed man crushed by a stationary robot at a Volkswagen plant in Germany, who died shortly after in a hospital (2015). Volkswagen claims that the death was due to human error–usually the robot is kept in a cage, but the worker, a third-party contractor, was working on the robot inside the cage. He was grabbed by the robot and pushed against a metal plate. I have been unable to find further news on this case, but Heron and Belford introduce their paper by citing it too (2015). A year earlier, Heron and Belford had published their *Scandal in Academia* case study, a compliment to the *Case of the Killer Robot*, and use this Volkswagen case for a discussion on where the blame should lie: the human, the hardware, the software, or a particular function call?

"[W]e can rarely," they say, "point to a single code point and say[,] 'That's the culprit.'" Just software alone is made up of layers upon layers of co-authored abstraction. They go on to cite a letter signed by big heads in computing, including Elon Musk and the now-late Stephen Hawking, calling for a ban on AI-powered weaponry; and they give the example that even when a human pulls the trigger in a drone strike, the technology of the drone itself is not too precise that civilians will not be caught in the fire. It is clear that Heron and Belford are arguing for robots not to be able to kill, because when they do, the blame is difficult to place–avoid that difficulty and we could all sleep easier at night. And while that would be a dream, they provide some comfort in the recognition that robot-mediated deaths are at least few in number compared to human-mediated ones. Still, that's not a perfect response. So they ultimately call for is a discussion: "[O]ur frameworks for having that discussion are not well equipped to deal with the logistics of distributed authority in software development... We can punish the human web around it, but we cannot truly punish an entity that has no conscious awareness of its own self."

Calo (*supra* Out) wrangles with how AI can increasingly predict the personal from the public (2017). This trend threatens consumers' ability "to appreciate the consequence of sharing information," and Fourth Amendment protections of reasonable expectations of privacy in public could go away. Calo compares *United States v. Jones* and *Florida v. Jardines* where, respectively: a GPS tracker affixed to a defendant's car was seen as a violation of that privacy without a warrant because the tracker also collected data while the car was at their home; and where a defendant lost a case involving a contraband-sniffing dog because the court ruled we have no reasonable expectation of privacy when it comes to contraband. More complicated, because AI is *predictive* and *automated*, AI does not constitute a search until a human lays eyes on the results, and even that can be filtered automatically to only the results where contraband or so on is predicted to be involved. Thus, Calo suggests, AI threatens to create a surveillance state where constitutional protections of privacy are side-stepped.

Angwin et al (2016b) and Larson et al (2016b) cover ProPublica's report on COMPAS. COMPAS is a proprietary software tool for calculating recidivism prediction scores. A higher score means a person, if released from prison, would be more likely to commit another crime within two years. COMPAS was used by a judge during sentencing, which falls outside the valid use for the COMPAS metrics, so this sentence was challenged in *Wisconsin v. Loomis*, though the challenge was unsuccessful (Rubel et al 2018). Angwin et al and Larson et al show than even within COMPAS's intended use it is biased against people of color. COMPAS scores, or tools like them, promise objectivity. However, their black box, algorithmic, and statistical nature threatens to undermine the constitutional right to individualized and equal justice.

Promote the Good Another goal of AI policy according to this literature is to promote AI and AI systems that have specific, fundamental, and good properties.

To give a few examples of these properties as argued for in the literature, Adams and Duarte hold that AI systems should align with human rights, function correctly, be resilient, not discriminate, provide redress, respect autonomy, be safe, be fair, and be explicable (2019). Baum et al (*supra* Pub) add trustworthiness, in particular a trustworthiness grounded in explainability (2019). Noothigattu et al add transparency and operation within social norms (2018b). And Yampolskiy and Fox, writing towards the discussion of ethics and super-intelligent AGIs, repeat the claims of safety and value alignment, and they add the technical restriction that these properties remain *provably* "even under recursive self-improvement," which, I add, we might say of any of the properties listed here (2013).

Summary 8.2 As suggested by this scholarly space, the goal of AI policy is to move us away from socio-technical systems with accountability gaps and systems that can be used to undermine fundamental constitutional and human rights; and towards an AI Ideal that maintains–in the face of its technological particulars and possible centralization of power–the inverse: clear and fair lines of blame when something goes wrong, and fair and just

protections of the fundamental rights that underpin a Good Life.

8.3 AI matters morally because regulating AI is difficult; the risks of AI are uncertain ex ante*; AI accountability lines are unclear* ex post*; although regulating AI is difficult, this difficulty is not insurmountable; and those tasked with, or those tasking themselves with, shaping AI policy have, or take upon themselves, a duty to surmount that difficulty*

The Limits of Codes of Ethics Several in this review believe that the AI industry cannot be trusted to hold itself accountable, by Codes of Ethics or otherwise.

Codes of Ethics are a form of policy under the general definition I'm taking here. And Codes are morally important for other reasons (generally *supra* Pro). But they fall short of fulfilling policy goals and may have the potential to directly undermine them: publics must take the company at its word; effectiveness is undermined by the disproportionate power held by just a few companies; and it does nothing to address misuse of AI technologies in open source projects[(xxxiii)] (CDT 2019). Codes that defer to AI law, solid on the surface, fall out from under themselves because the law on the matter is far from settled (EFF 2018b). The EFF challenges Google's 2018 AI ethics principles in this very vein: released after the employee outcry of Google's Project Maven controversy, the principles make no "commit[ment] to the type of independent, informed[,] and transparent review which would be ideal for ensuring the principles are always applied and applied well."

Privacy and the GDPR In order to protect and redress consumers with respect to data privacy, one argument is that AI policy may require attention to high-risk *inferences* in addition to its current protections.

Burton et al cites EU Article 22, which they see as an attempt to get ahead of AI's effects on fundamental rights, and which states:

> The data subject shall have the right not to be subject to a decision based solely on automated processing... Paragraph 1 shall not apply if the decision:... (c) is based on the data subject's explicit consent.

But Wachter and Mittelstadt take this "right to be forgotten" of the GDRP to task (2019). They argue for a "right to reasonable inferences" based on a "right on how to be seen." There is a gap, they say, between consumer's legally protected ability to rectify, delete, object to, and port *non*-inferred data compared to the counterpart, a lack of protections with respect to *inferred* data. When input data is inaccurate and the context of the inference is of high-risk, data subjects become harmed or wronged as a result of *inferred* data, and consumers may have no preemptive protection from these cases. Citing the Article 29 Working Party's guidelines (which are not legally binding, just provide a model), Wachter and Mittelstadt argue that because a high-risk inference is "likely to have an impact on a *certain* person's rights and interests" (emphasis added), then inferred data should be considered *personal* data. Therefore, they call for a two part solution. First, there should be an *ex ante* justification from data processers that the data to be inferred are reasonable and a justification that these are appropriate inferences to make based on the model's underpinnings. And second, there should be *ex post* mechanisms in place that would enable consumers to contest the inferences made.

The aims of law to provide privacy and discrimination protections, Wachter and Mittelstadt summarize, stemmed respectively from wanting to be left alone by the State and to protect culturally dangerous singling out following the horrors of WWII. These are important aims. However, the CDT (*supra* Pub) argue that privacy laws must be balanced against the ability to *assess* AI during research and audits[(xxxiv)] (2019). The CDT are not arguing that privacy laws should be foregone altogether, but that assessment of the safety and behavior[(xxxv)] of AI are priorities alongside privacy and non-discrimination too, and the specter of privacy-related infringement can dampen progress towards these.

Patents One concern is that overbroad patents for new AI technologies undermine AI policy's goals for innovation.

The EFF demonstrates this, showing how patent law can be used to fragment AI devel-

opment, such as Google's patent on the common "dropout" technique (2017b). Though this is an issue with software development in general, it does affect AI, and it can undermine other regulatory goals that depend on AI to actually progress. US Patent 5,944,839 "essentially covers using AI to diagnose computer problems" and US Patent 9,760,834 represents a trend feared by the EFF of patents "read[ing] like the table of contents of an intro to AI textbook... but applied to the specific example of [eg] proteins."

LAWS This literature holds that while Lethal Autonomous Weapon Systems (LAWS) could revolutionize warfare, humans must retain meaningful control over these systems.

Feldman et al's summary balances the strategic value of AI in war with its particular risks, such as an example of a man fooling an AI classifier by wearing an image, large and over his chest like a sandwich board, convincing the AI he is, instead, just another piece of landscape (2019). Humans will play some role in this AI war, so Feldman et al recommend: human-in-the-loop, where the AI acquires the lock but its the human that pulls the trigger, taking on the weight of responsibility; human-on-the-loop, where AI training scenarios involve humans in as real a situation as possible; and human-initiated, where AI is deployed to a bounded time and space, a necessary model when the speed of battle operations finally exceed human comprehension.

Compare these recommendations to DODD 3000.09 (Sayler 2019b). DODD 3000.09 requires that all autonomous weapons (lethal or not) be tested to ensure that they function as intended, that the AI will either halt its mission or confer with a human if it is unable to complete that mission on time, and that this intended functioning is robust against failures. DODD 3000.09 also requires that a human must be able to exercise judgement over any AI use of force, and that any changes to the weapon's AI models (such as through further ML training) will prompt a re-test of everything, old test results becoming invalid. Finally, DODD 3000.09 requires that any autonomous weapons with lethal capabilities undergo secondary, senior-level testing; this has never happened, according to Sayler, because

no autonomous weapons with lethal capabilities have been developed.

However, Calo is skeptical of any model that hinges on "meaningful human control" (2017). He argues that placing LAWS under human control does little besides absorbing liability from the AI. If a weapon system itself inherently challenges fundamental human rights, having a "fall guy" means it may be some time before the weapon system is decommissioned.

The Model of Tort Law Regulating AI like old technology may not be an appropriate choice of metaphor. One proposal is to use a government agency *ex ante* to set standards and certify AI systems and to use courts *ex post* to handle the complexities of the case at hand; only system owners with agency certification would face *limited* liability.

This is Scherer's proposal, "meant to start a conversation rather than to be the final word" and modeled on existing tort law (2016). Both AI and the technologies that precede it have public risks, but regulating AI the same way we have regulated older tech is difficult: AI's actions can be unexpected; control of AI behavior can be hard to maintain with precision; and it is difficult to determine "the 'who' and 'where' of potential sources of public risk." And while legislature lacks AI expertise, legislators can delegate policy-making to agencies staffed with experts. This approach, Scherer says, grew during the Progressive Era and again during the Great Depression, both faced with new issues of industrialization. However, agencies have been criticized as a "fourth branch" outside the standard democratic process, and this criticism was exacerbated by the New Deal, which created several new agencies. So, using agencies, Scherer says, is a "value choice." He envisions AI agencies that set *ex ante* industry standards to be coupled with *ex post* tort liability. Courts, he claims, are better equipped at handling the complexities of an industrial society where multiple actors may be involved in the harm. To address their limited *ex ante* powers, Scherer proposes an act which would create an agency whose role is to certify the safety of AI systems according to some future-facing technology standards. Those with certified systems would face *limited* liability if their system ever causes harms. Those without certification are conferred no such benefit. This,

he envisions, will place a heavier spectre on companies developing and deploying uncertified AI systems: either get certified, or spend greater resources in not causing harm (or in not getting caught).

Summary 8.3 This scholarly space presents a few themes for AI policy. A great balancing act must be made between the uncertainty of AI *ex ante* and the difficulty in determining accountability *ex post*. So, solutions frequently included auditing procedures and/or mixed approaches that required some certification or justification on the part of the AI company *ex ante* dove-tailed with protections for consumers *ex post*. So while AI poses legislative difficulties–this is certain–, these difficulties do not appear to be insurmountable.

There is, though, one more special set of considerations to discuss about AI policy, attended to in the next section: (when) should an AMA be considered a "person?"

9 Personhood

9.1 AI matters morally because treating AI with moral agency or moral patiency shapes AI policy in at least criminal liability, civil liability, and copyright law; treating AI with moral agency or moral patiency redefines humanistic intelligence into wholly technical terms; and text generated by AI today is convincing *that it has meaning or came from an intelligent source, so waiving away the matter of AI's personhood goes against one's best moral intuition*

> You will either be ascribing these marks [on the sand] to some agent capable of intentions or you will count them as the nonintentional effects of mechanical process. But in the second case–where the marks now seem to be accidents–will they still seem to be words? Clearly not. They will seem to *resemble* words. You will be amazed, perhaps, that such an astonishing coincidence could occur (Knapp and Michaels 1982).

I keep this quote posted above my desk as a reminder of the *amazement* we feel in AI. (Knapp and Michaels were writing originally about how to interpret poetry, looking at the hypothetical case of a poem that washes up on the beach.) AI can be amazing, but it cannot be a person. This is, at least, the humanist perspective. What it doesn't answer for us is

how we treat AI-the-actor in legal issues. This subsection takes up questions conceptually in this vein: questions around the personhood, agency, and patiency of AMAs.

What is an agent? The definition we use of a "moral agent" in the context of an AMA matters morally, technologically, and legally.

Generally speaking, a moral agent is able to respond to others, feel responsible for its own actions, and feel that others are responsible for their actions; one can view these criteria from the perspective of rational judgement or affective intuition; and one can view these criteria as binary or as scales (Williams n.d.). This is not an exhaustive account of moral agency as it has been taken up in Philosophy, but sufficient for grounding the works caught in this review.

Fabio Foss (in Poulsen et al 2019) provides a modified definition of moral agent for use in creating AMAs. He begins his argument by making it clear that the disciplines tasked today with creating AMAs *define* agency and morality, and the definition one works with of agency and morality guides their reasoning about what our moral relationship with AMAs should be. Within Computer Science, Fossa says, "agent" is a general actor in a system. This "applies both to human beings and thermostats." So, a "moral agent" is any general actor in a system that is capable of performing any action which can "cause moral good and evil" (Floridi and Sanders 2004, quoted in Poulsen et al 2019). Next, "moral reason" is the capability to select an action to perform, likely in response to a situation, *as though* that selection had moral meaning. Following this strict technical definition, Fossa rejects philosophical interpretations of AMAs that consider AMAs and humans as equivalent moral actors. Instead, Fossa argues for an interpretation of AMAs as "tools that execute functions autonomously. . . in accordance to given constraints that carry a moral weight." Our relationship with AMAs is obviously more complicated than ours with screwdrivers or automatic doors. In all three the moral responsibility is solely our own. But only with AMAs is there a heightened "moral sensitivity" as Fossa terms it.[xxxvi]

Persons in the law An AMA should not be treated the same as a human actor in criminal liability, civil liability, and copyright law, according to writers in this literature.

Lima looks at AI and criminal liability (2018). Criminal law requires an act, and that act requires agency. She gives the counterexamples of intoxication, impairments in reasoning faculties, unforeseeable consequences, sleep walking, and reflexes–all change the legal interpretation of the case. Criminal acts also require social relevance, such as breaking obligations, negligence, or undermining the rights of others. And third, criminal acts require *mens rea*: the knowledge that what one is doing is wrong. It is possible to *say* that AI fulfills *mens rea* because it could have been preprogrammed to know the law or social norms, but, Lima argues, this reduces choice, voluntariness, knowledge, and intent to just technical terms. We cannot permit this because it would undermine those values them as humanistic endeavors, and that would threaten to weaken how we perceive criminal law. On the other hand, not treating AI as having *mens rea* (or something similar) vastly expands the "scene of the crime" to include not only the victim and witnesses but also operators of the technology, the software, the hardware, the developers of the AI, project managers overseeing that development, and other AIs. But what would we gain, Lima asks, by holding AI itself as criminally liable? Punishing one AI is not likely to deter other AIs. Therefore, Lima holds that we emphatically should not treat AMAs as persons in criminal law because we should not weaken our perceptions of that serious area of law.

Civil liability law's function, Swanson says, is to protect consumers (2019). Swanson identifies an accountability gap in AI liability law because AI can be "used as intended but still cause an injury that neither the consumer nor the manufacturer foresaw." Swanson compares the laws of two states in this area. In New Jersey, when a consumer is harmed by an AI, the onus would be on that user under a claim of User Alteration, because the AI was "altered" at the hands of the user as a matter of its "learning" procedures. This is dangerous because it provides a shield to any AI company employing certain ML models. Conversely, in South Carolina, the onus is on the company because any AI outcome could

be argued to have been "foreseen" during R&D. And leaving the matter to the courts to decide (*supra* Pol, Scherer) will require large costs during litigation: "Current estimates for products liability expert witnesses in Washington run at approximately $250 per hour for initial case review and approximately $275 per hour for depositions and court appearance fees."

Palace argues that copyright for AI-generated works should go immediately to the Public Domain (2019). Copyright law requires that a work be original and fixed in a tangible medium. An AI-generated work is original and tangible. The question Palace explores, then, is not about the *work* but about *who* to give the copyright to. Giving the copyright to the AI itself has no legal standing he says. The purpose of copyright law is to protect the incentives of an artist to create. Considering this, he argues that while giving copyright to the programmers[14] would incentivize them to create, this would go too far and become an incentive to copyright anything their computer could possibly produce: "The creation of every painting [is not a] humanitarian goal." This leaves the Public Domain, which Palace argues still rewards the programmers: AI itself is generally lucrative in its own right, and the *code* behind the AI can still be copyrighted.(xxxvii)

Persons in service AMAs will one day be in service relationships with humans where the decisions made by any actor in that relationship are of special moral concern, such as in healthcare.

Coeckelbergh takes up the question, does a health care robot, stipulating that it must be moral, require a quality of "caring" in order to be moral (2010). Coeckelbergh answers "no." An AI having emotion(xxxviii) is not a necessary component for good AI-assisted health care, *even if* that is an assumption in the lay sense of "caring"–the operative move here is Coeckelbergh's focus on health *care* itself as the object of ethical inquiry, not the quality of *caring* in the actors that carry out that service. There are various arguments against the uncaring care robot, to which Coeckelbergh responds in turn.

[14]Or the business the programmers work for, or so on.

To the argument that care requires "an emotional exchange," Coeckelbergh says AI-assisted care may leave time for *more* human-human emotional exchange. And bluntly, "to be treated like a thing by a machine is less morally degrading than to be treated like a thing by a human being." To the argument that care with emotional exchange would be *better* care, he says responsibilities in healthcare are distributed, so AI-assistance can help distribute these requirements and result in a situation of care that is more holistically conducive of a good life for the patient. To the argument that Big Data AI undermines patient privacy, he asks, how private is a nurse washing me in a hospital, then says that privacy concerns are nothing new. Protecting privacy is a solvable problem. Lastly, to the argument that an AMA may fool the patient into *thinking* they were cared for, he imagines a "*eudemonia* machine." A *eudemonia* machine allows a physically sick patient to live an ideal *virtual* life. Someone in a *eudomonia* machine who never regains their potential, and someone who is never permitted the potential to regain that potential when the solution otherwise exists, violates Coeckelbergh's requirements for a good life. So AI care robots that operate in this way *do* violate Coeckelbergh's moral framework. However, a patient for which a *eudomonia* machine is used as a *means* to a good life is perfectly acceptable. I generally accept Coeckelbergh's rebuttals, except for the last (on fooling the patient into thinking they were cared for): I very much believe that a health care robot, as a technical matter, will cheat its reward function–say, satisfaction survey results–by behaving in a way that maximizes that signal and not in a way that is in the best interest for the patient; and so I would like to see work building on Coeckelbergh's last rebuttal that spells out the mechanism by which "reward cheating," so to say, could be detected and redressed.

Schwitzgebel and Garza (*supra* Per) add another caution to the idea of cheating (2015). A robot's AI "brain" might not fool us, but the design of the robot's *body* still can influence our behavior of how we treat it. They define this as the "Asimov Problem" (named after a performance by the ASIMOV robot one author witnessed). In this, a human finds a robot exceedingly cute and thus, hearing the robot and a human (both thrown overboard) scream,

saves the wrong one. The moral issue, I note, depends on where blame is placed. If I blame the robot, then the moral issue is that robots should not fool humans about their robot-status or be visually and aurally designed in a way that unfairly advantages itself over a human, based solely on humans' sensory perceptions of the device. But if I place the blame on the human doing the saving, then the moral issue is that a human should not value the live of a replaceable non-human over an irreplaceable human, such as the case of a human saving their long time robot butler over a human stranger.(xxxix)

What if we treat AMAs as moral agents? It is gross and cruel to murder an AMA; however, that is not ripe[15] because we are a ways off from creating an artificial being we could "murder"; and creating AMAs and treating them as having moral agent status, some in this review worry, may lead to a "utility monster," ie, a being that greedily exceeds its fair share of resources in no human's best interest.

Schwitzgebel and Garza proceed from a "psycho-social" definition of moral status, then claim that sufficiently advanced AI will have no relevant psychological or social difference from humans (2015). From this starting point they explore two branches of thought: First, if it *is* possible to create AI of this kind, then it is possible to create a "utility monster." They imagine a society with an allocation of cookie resources. If cookies are allocated equally to all members of the society, then an AI utility monster could replicate itself (creating thousands of clones), collect a large distributive share of cookies (thousands of cookies instead of the original allocation of just one), then re-combine into one moral entity who is now in possession of too many cookies.(xl) Second, if the issue of replication and re-combination is solved, then AI moral agents would have, Schwitzgebel and Garza argue, greater moral status than we attribute to a human *stranger*.(xli)

They imagine that one could make a religious rebuttal to treating AMAs as moral agents because an AI lacks a soul. Even so, they respond, there is no relevant difference between

[15]In *United Public Workers v. Mitchell*, the Supreme Court "refused to enjoin a merely 'hypothetical threat' because it presented no actual case or controversy," ie, the case was not ripe because it had not happened yet (Hall, Kermit. *The Oxford Companion to the Supreme Court of the United States.* 2005.)

an AI and a human iteratively replaced with artificial limbs. However, to kill a robot is less immoral intuitively than killing a human. Yet, it is immoral to demote the moral status of an entity once you've discovered it is a machine, and it is cruel to recklessly create life just to be destroyed. But, even if it would be cruel and gross, it would still be no less legal than the *simulation* of child pornography by consenting adults.

What if we treat AMAs as moral patients? One argument is that AI, if sufficiently human-like and/or promising to let "humanity" live on forever even after the extinction of humans, should be given higher moral patiency status than everyday possessions; however, it may still be more ethical to not build human-like AI that competes with us for resources.

Bryson takes this view of AI moral patiency (2018). Bryson asks, should AMAs be given moral protections higher than we already assign to ordinary possessions? Bryson gives the examples of the Bible, Koran, the American Flag, and children.[(xlii)] Children, however, are treated with moral patiency in order to protect them so that they can grow up and become adults themselves, and at that point have full moral agent status; AMAs don't "grow up" in this same sense, and with that she rebuts the child analogy. As a possession of higher interest, Bryson does recognize the comparison to Kant's "animals," where treating a human-like entity as less than human will only lead us to de-humanize other humans; and she recognizes that treating AI as moral agents (the topic of the selection above) simplifies legal matters in a system that grows ever more complex and opaque. However, she responds, don't make opaque systems, don't make human-like robots, and don't build things that will compete with us for resources. And even though the Ideal of building an immortal being gives us a promise that humanity, in some form, will live on forever, Bryson concludes that the benefits do not outweigh the issues, and so AI should not be given higher moral patiency.

Summary 9.1 So, in sum, this scholarly space is asking the "hard questions" around, for example, how we should treat intelligent beings, what it means to give an artifact intelligence, how that act will affect how we define intelligence in ourselves, and how to regulate new

technology in general when facing a growing accountability gap. At one level, these are questions about "authorship": what is it that makes human text, human opinion, human feeling–*meaningful*? How we answer this affects how we answer questions such as around interpreting AI output and around copyright. At a higher level, these questions require us to look at the goals of using AI and AI policy in human endeavors. How does the introduction of a new "agent" or "patient" potentially upset those endeavors? In questions around harms and wrongs, a recurring issue is that of accountability gaps, where the upset isn't (just) that there is a new "agent," but that the (clear lines to) human agents we are used to holding morally accountable have been removed or obscured. A look across the literature of this section shows that there are *several* metaphors being deployed, and there does not appear to be one metaphor for AI in policy that always "wins out" and gives us all the answers. This is why the questions of this space are "hard questions": figuring out how different ways of conceiving AI makes some issues clearer and others more complicated is asking us to dig down deeper to the distinctions–the morally salient features of the situation–that matter.

10 Conclusion

The purpose of this literature review was to provide a canvas of conceptions on AI as a moral matter. Whereas other canvases found in the review generally took a taxonomization approach, I took the approach of an exhaustive (til convergence) review to gain a sense of the broader discursive formation around AI as a moral matter. I then presented the facets of that formation in nine conception headings, one excluded (epistemic trust) to be the subject of a future paper.

I know that there are limitations to my approach (see the Appendix on this). But I believe that this adds to the literature a roadmap: not a roadmap of where to go next, but a roadmap of roads not taken. Each way we conceive of AI leads us to different moral matters, different questions, different rhetorical movements, and different conclusions. This conception of the literature is also useful to me for my own teaching practice: I do not know

what background students will be coming from or what prior experiences with AI they will bring to the classroom. This varied conception of the discursive formation, then, provides me invaluable experience with appreciating and canvasing together such wide thought.

This varied conception also shows that for we teachers, *whose* AI we teach is a moral choice:

- The situation we will face in CDIS will be an *interdisciplinary* one, so this will not always be an "easy" choice where we can fall back on disciplinary convention.
- We have a mission to outreach to the community, in part under the Wisconsin Idea, and so once we do outreach to the community, we will also be reproducing the *perception* of the AI/Data Science profession to those outside of it.
- Many of our students will go on to be AI practitioners, but not all. Some may become AI policymakers or so on. Simply put, AI will touch nearly every sector. And while we cannot possibly teach students who will enter every sector, we may still find ourselves collaborating with partners that *do* reach those areas. We may also find ourselves training government-funded, non-traditional students as part of a scholarship for service or similar program meant to mitigate the displacement effects of AI on current workers. Regardless, *all* our students–with diverse backgrounds and diverse goals–will need to learn the downstream societal consequences that can arise from their decisions due, for example, to biased data, mediation of human discourse, and special sensitivity in institutions of high scrutiny.
- Students working on self-driven AI capstone projects, whether in a class under our department or under one of our partners, will likely be drawn towards the kinds of AI interactions that they are already familiar with, and students will be incentivized to do just that kind of work: employers may be expecting examples of current AI trends in the student's portfolio. Those familiar AI interactions have side-effects not obvious at the time the student is coding. Will we provide events, for example, where students can demonstrate and test their AI projects with others from backgrounds different

than their own so that errors like the "Taiwanese eyes" false positives will be caught and discussed in a pedagogic, productive environment? I believe we should. And I also believe that we should take seriously students' autonomy to decide for themselves what projects to pursue. But to do this we must be prepared to *do real work* to give their degree program the support they will need to (1) do well on those projects in the short-term and (2) model through the doing of those projects how to do Good work later in life.

- Because we will likely be teaching AI Policy as part of our larger endeavor to provide AI Ethics support to our partners under CDIS, we need to be cognizant of our own role in the broader "AI Standardization" process, as one author in this review named the complex socio-technical landscape of how AI Policy comes into being in multiple contexts and through competing interests. So, our teaching of AI Policy will include modeling for our students how to engage with AI Standardization both now and throughout their professions. Put another way, we have an opportunity to model for the students *real* policy engagements. We teachers at UW-Madison are quite nearby the state capital: just a twenty minute walk at a slow pace. Government agencies dealing with AI, data, or so on–these likely would be happy to host discussions and/or presentations with small classes of students.[xliii] Students might also enjoy these encounters, and putting a face to policy-making may help them to return down the road to thinking, in a well-informed way, about AI Policy and the policy implications of their work. And so the choice *to* model or *how to* model policy engagement is one with very possible downstream consequences.

This conclusion, synthesizing the views found in the literature on AI as a moral matter and directed towards our present situation here, only further motivates me to answer my bigger question, "How should we teach AI Ethics?" All of this sets up a lot of ideas of future work–including the two planned papers on (1) AI/Epistemology and (2) Our Moral Responsibilities as a University–, way more than I could possibly do in one academic lifetime.

Still, whatever that future work may be, and whatever I may have missed in this broad review that I will pick up along the way, I am excited for it.

Notes

(i) Example endnote

(ii) I am thinking of Justus Randolph's "A Guide to Writing the Dissertation Literature Review" with these definitions (2009).

(iii) I am thinking of Foucault.

(iv) Burton et al (*supra* Out) point out that AI has been used in fiction as a narrative device, sometimes focusing on an AI that should not be trusted, and at other times on the human developer (2017). This reference to fiction still treats AI as a narrow symbol. While restricting the AI to an algorithm, statistical technique, or robot tasked with some stated goal can be helpful in cases, it ignores art's wider aim to say, through the mirror of that which is different from ourselves, something that is *very* about ourselves. This too is a valuable human endeavor.

(v) I am thinking of CU-Boulder's Casey Fiesler, the creator of the extensive "Fiesler list" of AI Ethics courses.

(vi) I am thinking of "epistemic injustice" or "discredited knowledge" on this last point here.

(vii) See my MA thesis, *Diversity, Identification, and Rhetoric in Tech*, for a more at length application of Burke to the tech industry.

(viii) See also Sharma et al, who list our social duties as responsibly designing, deploying, and maintaining AI models (2019). Why these stages (design/deployment/maintenance)? I believe that Sharma et al are relying on an appeal that would go like this: The responsibilities of our profession do not stop at the enter key, but continue through the iterative design process of the software development life cycle (SDLC) *as well as* the ongoing effects on society our profession has *and* the effect society's response to our technology has on our conceptions of values during that SDLC. This is similar to the interpretative approach used by Balsamo to remove the technology/society false dichotomy (2011).

(ix) I believe "we want to develop curriculum" needs emphasis. A full treatment on the interrelated goals of instructors and the goals of practitioners, especially in Computer Science where increasingly teachers hold dual appointments with Academia and Industry, is outside the scope of this paper–though it's on my list. I also can't give a full treatment here on the interrelated effects of curricula and Codes of Ethics on the identity of the profession. A future paper is planned on why we *morally should* want to develop such curricula.

(x) I am thinking of course of the debates I studied around the Contributor Covenant. A discussion of this is given in my MA thesis.

(xi) See also my MA thesis on Carlyle's "clothes" brought into the Burke discussion.

(xii) See also Katie Shilton's work on value levers, which takes a Good Life ethics approach and looks at how day-to-day practices of developers affects their ethics deliberating.

(xiii) Getting a little ways from consumers, investors and donors need trust too, trust that AI initiatives they fund, or public initiatives they fund with possible AI components, will return on their investments, in some cases a financial return and in others a moral return. Thomas and Ornstein (2018) and Ornstein and Thomas (2018a) make this claim in tracking the Memorial Sloan Kettering Cancer Center (Sloan) case. In this case, conflicts of interest threaten donor's trust that charity funds will return on the mission to use AI to improve cancer treatments, which in turn threatens fundraising initiatives that staff need to carry out that mission. The conflict of interest comes in executives who received stock options from a now-public biotech company. It is important that execs have close ties with industry actors in order to make the kinds of medical breakthroughs that excite donors, but because Sloan is nonprofit, execs also should not received personal compensations from these industry deals. Pathologists at Sloan "complained that their work was being commercialized for private gain" and expressed worries that private patient data was being shared with third parties either without patient consent or without proper communications. Sloan has responded with stricter policies on disclosing industry ties; still, this case illustrates the need for a systemic assessment argued for above.

(xiv) I wrote this before news of the fatal Uber crash

(xv) Analyzing this, if AI is not accurate or is unfair against me, then it goes against my best interest to trust it or those who support it. If AI is unfair towards those unlike me, yet I benefit from it, then it may be in my best short-term financial (or similar) interest, but it will not be in my best existential interest: because I believe (a future paper is planned on this) knowledge is transmitted in a co-agent way through the community, and because I believe to say "I know" requires an honest due diligence, then I lose sincerity and ability *to even know about the world* any time I knowingly support an AI application that unfairly treats those unlike me. That type of unfairness over time excludes those unlike me from my community and discredits their knowledge. Similar arguments could be made about a government project on AI that has not been transparent, or a government AI technology that has been applied in an inappropriate context. However, no government AI project can be perfectly fair, transparent, and applied on the first try. And it *might* improve with successive revisions. Struggling to define what the inverse of "trust" is raises three questions: When should we *contest* an AI that we distrust, when should we *ignore* it, and when should we *support its revision*? I'll say for now, context matters. See also Coeckelbergh (*supra* Mor) who raises a different concern that supports the CDT's recommendations (2010). Writing about AI deceit, Coeckelbergh leads me to imagine a human who does not know they are interacting with an AI, or a human who incorrectly misplaces the blame for some harm onto some other (non-AI) entity or behavior in the AI system. And see also Poulsen et al (2019), Buechner (2013), Burnett et al (2013), and Baum (2019) on trust as a fundamental requirement of machine-embedded ethics and of developers' good conscience to deploy their technology.

(xvi) Poulsen et al is a paper by eight authors (2019). It is a dialog between these authors as they collectively respond to a paper by van Wynsberghe and Robbins (2018). The authors in Poulsen et al respond to these critiques in turn, each author penning a response, and they often disagree with one another. This is. . . weird. . . in comparison to the other papers caught in this review. Adam Poulsen stumbled across an earlier draft of this review and emailed me to respond to this "weird" comment: "You are so right! I had a good laugh at myself. Although it is an unorthodox manuscript, I hope it was somewhat valuable in informing readers on the true nature of the debate around this diverse, cross-disciplinary topic." I let him know how helpful his work was for this project, and I've since put Adam's email in my Outlook folder for emails that make me smile.

(xvii) Again, I wrote this before the Uber news.

(xviii) Such as, now I guess, putting into words that, yes, a jaywalker is still a person.

(xix) I owe this point to Alan.

(xx) One paper reverses the arrow, troublingly, to reimagine humans as biological machines lacking free will, and thus arguing for a strict preference-utilitarian system of ethics (Klein 2015). His concluding motivation is that when one day AI has gained consciousness and the bots around us wake up, then we must hope that our ethics works for both the human and AI species. His own hope is that he has moved the needle in that direction; however, I cannot take this piece seriously when an underlying assumption is the sticky and specious interpretation of the human mind as "machinery." See also Vanderelst and Winfield, who caution that creating ethical machines creates–not like a shadow as the language suggests–unethical ones too (2016). These unethical machines can emerge when a single robot learns to hack its reward function, an ethical robot suddenly behaving in unanticipated ways. They can emerge when competing interests among a network of AMAs leads to unethical *group* behavior. And they can emerge when code is modified *very* slightly, either due to a bug, a bad software update, or a malicious hack. They give the example of a robot assisting a human playing a shell game against another human. Should the robot, seeking to help its human win, motion in such a way that the opponent (1) trusts the robot and (2) takes the robot's advice to pick up the wrong shell? I can imagine scenarios where either answer would intuitively seem correct. And the hinge between these scenarios is what type of moral relationship we want to have with machines, beyond merely asking about agency and patiency: when the human players are young children, certainly the robot should not help its human cheat, because we do not want our children to model that behavior; but when the humans are adults competing (I have seen it argued that a discussion board simulating slavery is legal so long as only consenting adults are involved; applied here, this might go: I have the right as an adult to consent to play against a cheating robot), then the issue goes away.

(xxi) "Human affairs": I am thinking of Herbert Simon's work, which was a heavy influence on me during undergrad.

(xxii) I owe this point to Alan

(xxiii) See also Delgrande et al (2018) and McCarthy (2003) on these points.

(xxiv) EPS 870 is a semester-long *part one* on these theories alone.

(xxv) At the time I am revising this, Juul has since pulled their more fruity flavors from shelves.

(xxvi) Unsurprisingly, but still sad, at the time of revising there has been another that has made national news.

(xxvii) Again, the theory on this is *vast*.

(xxviii) I am thinking of Burke's "mystical" and "metonym."

(xxix) If AI *does* replace jobs, generally speaking, the idea has been floated to tax it, an idea agreed with by Bill Gates (Calo 2017).

(xxx) Consider too Yoshua Bengio's, "A lot of what is most concerning is not happening in broad daylight" (interview in Castelvecchi 2019).

(xxxi) The papers captured in this review almost certainly all stem from Western sources. A better comparison of policy on AI across these cultural clusters is outside the timeline for this paper and outside my own linguistic toolkit.

(xxxii) As Nature points out in the example of facial recognition companies turning down partnerships (eg, with police) when they view the application as immoral, when it's left to companies to make these types of decisions, then it is left to the very few to make moral judgements in the best interest of all (2019).

(xxxiii) See also generally the Contributor's Convenant which seeks to establish a Code of Ethics for open source projects. While technically opt-in, it is the default option for one popular open source project management tool (Knowles 2018).

(xxxiv) The CDT's model suggests audits from within the developing company and audits by third-party auditors. The audit itself should cover the entire socio-technical AI system and include checks on how malaffected users will be addressed.

(xxxv) See also Sharma et al on an assessment framework based on counterfactual burden scores (2019); and Calo for a discussion of certification of AI systems, compared to a pilot's flight school (2017)

(xxxvi) Alan asks, "Where's the issue of human agents *using* AI and what is their responsibility?" See generally *supra* Mor on AMAs aiding human moral decision making and *supra* Pol on meaningful human control and related questions.

(xxxvii) I am not sure though how cleanly these lines would be cut in, for example, the work of Joanne Hastie.

(xxxviii) See also Lawrence et al, which develops an argument that ethical AI, if it is to be recognizable by us, will require a shared nature, evolution, social history, and/or political history (2016). They arrive at this conclusion after an exploration of the *foundations* of human moral action–making no claim about what makes up an ontological essential component of moral action. These foundations are the human concepts of passion, vulnerability, reason, justice, and reciprocity. In other words, human moral action is founded on what is *felt*; and so a view of human/AI interaction that does not consider the felt effects of AMAs on our own morals is a limited view.

(xxxix) See also Etzioni and Etzioni, who describe an interesting relationship between passenger and their self-driving car (2017). They proceed from a definition of AMAs as *only* acting ethically with relation to what is left undefined by law. It is not a moral choice to follow the speed limit or stop at a stop sign if one has already accepted that use of AMAs requires public adoption, which informs law, which represents a publicly agreed upon constraint on machine behavior. But whether the self-driving car stops for hitchhikers or drives at a slower speed that also reduces pollution–these are moral decisions for the car. However, as Etzioni and Etzioni point out, evidence in HCI repeatedly points out that humans tend to accept the default configuration for technology. So those moral decisions would, implicitly for the general use of the vehicle, be set by the developers. To propose an architecture that solves this issue, the authors conjure up the image of an "ethics bot," which they define as an AI that analyzes the behavior of some particular individual (say, its owner) in order to construct a model of that person's moral preferences–with respect to the behaviors left undefined by law. They point to Nest's smart thermostat as an existing example, if the temperature of one's house is taken to be a moral choice (which I generally do not agree with). I do find it interesting how Etzioni and Etzioni's conception of AMAs as learning moral preferences within society-defined constraints solves the tension between Anderson and Anderson and Fossa (in Poulsen et al 2019), while raising a new question: I can appeal to law in moral "code" to remove the paternalistic relationship between developers and users, yet is individual moral preference, as some gradient within the confines of law, something "optimizable?" Does sliding our configuration along that scale, finding "just the right spot" so to say, actually lead to morally better behavior?

(xl) See also Englert et al for a discussion of allocation of rights to AI to public spaces like sidewalks and roadways (2014).

(xli) See also Scheessele for an argument that AI's moral status will be that of plants (2018); and Estrada for an argument that we treat AI as moral agents in professional pursuit of a principles, though he recognizes this area of law stemmed from slavery (2018).

(xlii) See also Giuffrida et al's example of treating AI as a child or an adult with Savant Syndrome (2018)

(xliii) Compare to ELPA 765: Education Policy Analysis, which does this with the Department of Public Instruction.

Bibliography

Adamopoulous, Panagiotis and Alexander Tuzhilin. "On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected." *TIST* 5(4). 2014. A2.24.

Adams, Stan and Natasha Duarte. "Just be Ethical: High Level Guidelines on AI are Fine but Offer Little Guidance." *CDT*. 2019. B1.1.

Aghaei, Sina, et al. "Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making." *arXiv Preprints*. 2019. A1.9.

Angwin, Julia and Jeff Larson. "Machine Bias: Bias in Criminal Risk Scores is Mathematically Inevitable, Researchers Say." *ProPublica*. 2016. B5.7.

Angwin, Julia, et al. "Breaking the Black Box: When Machines Learn by Experimenting on us." *PrpPublica*. 2016a. B5.6.

Angwin, Julia, et al. "There's Software used across the Country to Predict Future Criminals and it's Biased against Blacks." *ProPublica*. 2016b. B5.8.

Arnold, Thomas and Matthias Scheutz. "Beyond Moral Dilemmas: Exploring the Ethical Landscape in HRI." *HRI*. 2017. A2.7.

Awad, Edmond, et al. "The Moral Machine Experiment." *Nature* 563. 2018. B3.2.

Baek, Jongmin Jerome. "How to Solve Moral Conundrums with Computability Theory." *arXiv Preprints*. 2018. A1.6.

Balsamo, Anne. *Designing Culture: The Technological Imagination at Work*. Duke. 2011.

Baum, Kevin, et al. "Towards a Framework Combining Machine Ethics and Machine Explainability." *arXiv Preprints*. 2019. A1.8.

Baum, Seth. "Social Choice Ethics in Artificial Intelligence." *AI & Soc*. 2017. B4.15.

Beckers, Sanders. "AAAI: An Argument Against Artificial Intelligence." n.d. B4.12.

Bjork, Ingrid and Iordanis Kavathatzopoulos. "Robots, Ethics, and Language." *SIGCAS* 45(3). 2015. A2.16.

Bosco, Dominic. "Standards for Artificial Intelligence can Shape a More Secure and Equitable Future." *CDT*. 2019. B1.2.

Brighouse, Harry. *On Education*. Routledge. 2006.

Bryson, Joanna. "Patiency is not a Virtue: The Design of Intelligent Systems and Systems of Ethics." *Ethics and Information Technology* 20. 2018. B4.1.

Buechner, Jeff. "Trust and Ecological Rationality in a Computing Context." *SIGCAS* 43(1). 2013. A2.17.

Burke, Kenneth. *A Rhetoric of Motives*. University of California Press. 1969.

Burnett, Chris, et al. "Stereotypical Trust and Bias in Dynamic Multiagent Systems." *TIST* 4(2). 2013. A2.25.

Burton, Emanuelle, et al. "Ethical Considerations in Artificial Intelligence Courses." *arXiv Preprints*. 2017. A1.4.

Calo, Ryan. "Artificial Intelligence Policy: A Primer and Roadmap." 2017. B6.4.

Castelvecchi, Davide. "AI Pioneer: 'The Dangers of Abuse are very Real."' *Nature News Q&A*. 2019. B3.4.

CDT. "Comments of the Center for Democracy & Technology on European Commission's High Level Expert Group on Artificial Intelligence (AI HLEG)'s Draft Ethics Guidelines for Trustworthy AI." *CDT*. 2019. B1.3.

Chopra, Amit K. and Munindar P. Singh. "Sociotechnical Systems and Ethics in the Large."

AIES. 2018. A2.8.
Chopra, Amit K., et al. "Research Directions in Agent Communication." *TIST* 4(2). 2013. A2.30.
Coeckelbergh, Mark. "Health Care, Capabilities, and AI Assistive Technologies." *Ethic Theory and Moral Practice* 13. 2010. B4.2.
Coeckelbergh, Mark. "The Invisible Robots of Global Finance: Making Visible Machines, People, and Places." *SIGCAS* 45(3). 2015. A2.14.
Correll, Michael. "Ethical Dimensions of Visualization Research." *arXiv Preprints*. 2018. A1.10.
Danaher, John. "The Rise of the Robots and the Crisis of Moral Patiency." n.d. B4.13.
Delacroix, Sylvie. "Taking Turing by Surprise? Designing 'Digital Computers' for Morally-Loaded Contexts." *arXiv Preprints*. 2018. A1.3.
Delgrande, James, et al. "General Belief Revision." *JACM* 65(5). 2018. A2.1.
Dignum, Virginia, et al. "Ethics by Design: Necessity or Curse?" *AIES*. 2018. A2.5.
Dockterman, Eliana. "Robot Kills Man at Volkswagen Plant." *Time*. 2015.
Drozdek, Adam. "Moral Dimension of Man and Artificial Intelligence." *Open Forum*. n.d. B4.14.
Durkheim, Emile. *The Division of Labor in Society*. The Free Press. 1985.
Edmonds, David. *Would You Kill the Fat Man?* Princeton. 2014.
EFF. "Artificial Intelligence & Machine Learning." *EFF*. n.d. B1.18.
EFF. "Google should not Help the U.S. Military Build Unaccountable AI Systems." *EFF*. 2018a. B1.13.
EFF. "Help EFF Track the Progress of AI and Machine Learning." *EFF*. 2017a. B1.16.
EFF. "How Good are Google's New AI Ethics Principles?" *EFF*. 2018b. B1.19.
EFF. "How Militaries should Plan for AI." *EFF*. 2018c. B1.20.
EFF. "Stupid Patent of the Month. Will Patents Slow Artificial Intelligence?" *EFF*. 2017b. B1.17.
EFF. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." *EFF*. 2018d. B1.14.
Englert, Matthias, et al. "Logical Limitations to Machine Ethics." *arXiv Preprints*. 2014. A1.5.
Estrada, Daniel. "Value Alignment, Fair Play, and the Rights of Service Robots." *arXiv Preprints*. 2018. A1.7.
Etzioni, Amitai and Oren Etzioni. "Incorporating Ethics into Artificial Intelligence." *J Ethics* 21. 2017. B4.10.
Ezrachi, Ariel and Maurice E. Stucke. "Artificial Intelligence & Collusion: When Computers Inhibit Competition." 2017. B6.2.
Feldman, Philip, et al. "Integrating Artificial Intelligence into Weapon Systems." *arXiv Preprints*. 2019. A1.11.
Garcerek, Ursula and Detlef Steuer. "Approaching Ethical Guidelines for Data Scientists." *arXiv Preprints*. 2019. A1.12.
Gerdes, Anne. "The Issue of Moral Consideration in Robot Ethics." *SIGCAS* 45(3). 2015. A2.18.
Gillum, Jack and Jeff Kao. "Aggression Detectors: The Unproven, Invasive Surveillance Technology Schools are using to Monitor Students." *ProPublica*. 2019. B5.5.

Guifridda, Iria, et al. "A Legal Perspective on the Trials and Tribulations of AI: How Artificial Intelligence, the Internet of Things, Smart Contracts, and Other Technologies will Affect the Law." *Case Western Law Review* 68(3). 2018. B2.7.

Gunkel, David J. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT. 2012.

Harris, Laurie. "Overview of Artificial Intelligence." *CRS*. 2017. B1.10.

Heron, Michael James and Pauline Belford. "Ethics in Context: Scandal in Academia." *SIGCAS* 44(2). 2014. A2.15.

Heron, Michael James and Pauline Belford. "Fuzzy Ethics: Or How I Learned to Stop Worrying and Love the Bot." *SIGCAS* 45(4). 2015. A2.19.

Jentzsch, Sophie, et al. "Semantics Derived Automatically from Language Corpora Contain Human-Like Moral Choices." *AIES*. 2019. A2.9.

Jeppsesen, Jens-Henrik, et al. "EU Tech Policy Brief: July 2019 Recap." *CDT*. 2019. B1.4.

Kasenberg, D., et al. "Quasi-Dilemmas for Artificial Moral Agents." *arXiv Preprints*. 2018. A1.13.

Kim, Tae Wan, et al. "Grounding Value Alignment with Ethical Principles." *arXiv Preprints*. 2019. A1.14.

Klein, Wilhelm. "Robots make Ethics Honest–and Vice Versa." *SIGCAS* 45(3). 2015. A2.20.

Knapp, Steven and Walter Benn Michaels. "Against Theory." *Critical Inquiry* 8(4). 1982.

Knowles, Bryan. "Coraline Ada Ehmke: Promoting Richer Open Source Communities." *XRDS* 24(3). 2018.

Larson, Jeff, et al. "Breaking the Black Box: How Machines Learn to be Racist." *ProPublica*. 2016a. B5.4.

Larson, Jeff, et al. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*. 2016b. B5.10.

Lawrence, David, et al. "Artificial Intelligence: The Shylock Syndrome." *Cambridge Quarterly of Healthcare Ethics* 25(2). 2016. B2.8.

Lee, Sau-lai and Ivy Yee-man Lau. "Hitting a Robot vs. Hitting a Human: Is it the Same?" *HRI*. 2011. A2.6.

Lima, Dafni. "Could AI Agents be Held Criminally Liable: Artificial Intelligence and the Challenges for Criminal Law." *South Carolina Law Review* 69(3). 2018. B2.3.

Lu, Joyce and Laurie Harris. "Artificial Intelligence (AI) and Education." *CRS*. 2018. B2.2.

Lucas, Richard. "Moral Theories for Autonomous Software Agents." *SIGCAS* 34(1). 2004. A2.21.

Malle, Bertram F., et al. "Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents." *HRI*. 2015. A2.11.

Malle, Bertram F., et al. "Which Robot am I Thinking About? The Impact of Action and Appearance on People's Evaluations of a Moral Robot." *IEEE*. 2016. A2.12.

Margetts, Helen and Cosmina Dorobantu. "Rethink Government with AI." *Nature* 568. 2019. B3.5.

McCarthy, John. "Problems and Projections in CS for the Next 49 Years." *JACM* 50(1). 2003. A2.2.

McElfresh, Duncan. "A Framework for Technically- and Morally-Sound AI." *AIES*. 2019. A2.29.

Mittelstadt, Brent, et al. "The Ethics of Algorithms: Mapping the Debate." *Big Data and Society*. 2016.

Muller, Vincent C. *Fundamental Issues of Artificial Intelligence*. Springer. 2016.

Nath, Rajakishore and Vineet Sahu. "The Problem of Machine Ethics in Artificial Intelligence." *AI & Soc*. 2017. B4.8.

Nature. "Time to Face Decisions." *Nature Editorial*. 2019. B3.7.

Noothigattu, Ritesh, et al. "A Voting-Based System for Ethical Decision Making." *arXiv Preprints*. 2018a. A1.2.

Noothigattu, Ritesh, et al. "Interpretable Multi-Objective Reinforcement Learning through Policy Orchestration." *arXiv Preprints*. 2018b. A1.16.

Nunez, Catherine. "Artificial Intelligence and Legal Ethics: Whether AI Lawyers can make Ethical Decisions." *Tulane Journal of Technology and Intellectual Property* 20. 2017. B2.1.

Oesterheld, Caspar. "Formalizing Preference Utilitarianism in Physical World Models." *arXiv Preprints*. 2015. A1.18.

Ornstein, Charles and Katie Thomas. "Cancer Center Switches Focus on Fundraising as Problems Mount." *ProPublica*. 2018a. B5.3.

Ornstein, Charles and Katie Thomas. "Sloan Ketting's Cozy Deal with Start-Up Ignites a New Uproar." *ProPublica*. 2018b. B5.9.

Overdorf, Rebekah, et al. "POTs: Protective Optimization technologies." *arXiv Preprints*. 2018. A1.15.

Palace, Victor. "What if Artificial Intelligence Write This: Artificial Intelligence and Copyright Law." *Florida Law Review* 71(1). 2019. B2.4.

Park, Yubin and Joydeep Ghosh. "CUDIA: Probabilistic Cross-LEvel Imputation using Individual Auxiliary Information." *TIST* 4(4). 2013. A2.27.

Pavaloiu, Alice and Utku Kose. "Ethical Artificial Intelligence, An Open Question." *arXiv Preprints*. 2017. A1.17.

Peng, Wei, et al. "Mining the 'Voice of the Customer' for Business Prioritization." *TIST* 3(2). 2012. A2.26.

Petit, Nicolas. "Law and Regulation of Artificial Intelligence and Robots: Conceptual Framework and Normative Implications." 2017. B6.3.

Poulsen, Adam, et al. "Responses to a Critique of Artificial Moral Agents." *arXiv Preprints*. 2019. A1.1.

Rahwen, Iyad, et al. "Machine Behavior." *Nature* 568. 2019. B3.1.

Raso, Filippo, et al. "Artificial Intelligence & Human Rights: Opportunities & Risks." *Berkman Klein Center*. 2018.

Reddy, Raj. "Three Open Problems in AI." *JACM* 50(1). 2003. A2.4.

Rodriguez, Marko, et al. "Faith in the Algorithm, Part 2: Computational Eudaemonics." *arXiv Preprints*. 2009. A1.19

Rubel, Alan et al. "Algorithms, Agency, and Respect for Persons." 2018.

Russel, Stuart and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson. 2009.

Samaniego, Jose Miguel. "Practices of Governing and Making Artificial Intelligences." *Disputatio* 7(8). 2018. B4.6

Santos-Lang, Christopher. "Our Responsibility to Manage Evaluative Diversity." *SIGCAS*

44(2). 2014. A2.22.
Sarwar, Shihab, et al. "Physician Perspectives on Integration of Artificial Intelligence into Diagnostic Pathology." *NPJ Digital Medicine* 28(2). 2019. B3.3.
Sayler, Kelley. "Defense Primer: Emerging Technologies." *CRS*. 2019a. B1.12.
Sayler, Kelley. "Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems." *CRS*. 2019b. B1.11.
Scheessele, Michael. "A Framework for Grounding the Moral Status of Intelligent Machines." *AIES*. 2018. A2.10.
Scherer, Matthew. "AI in HR: Civil Rights Implications of Employers' Use of Artificial Intelligence." *SciTech Lawyer* 13(2). 2017. B2.6.
Scherer, Matthew. "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." *Harvard Journal of Law & Technology*. 29(2). 2016.
Schwitzgebel, Eric and Mara Garza. "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* 39. 2015. B4.4.
Segal, Michael. "Q&A: Andy Clark: A Philosopher's View of Robots." *Nature*. n.d. B3.6.
Sharma, Shubham, et al. "CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence." *arXiv Preprints*. 2019. A1.21.
Shaw, Nolan P. "Towards Provably Moral AI Agents in Bottom-Up Learning Frameworks." *AIES*. 2018. A2.13.
Surden, Harry. "Machine Learning and Law." *Washington Law Review* 89. 2014. B6.1.
Susser, Daniel. "Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures." 2019. B4.9.
Swanson, Greg. "Non-Autonomous Artificial Intelligence Programs and Products Liability: How New AI Products Challenge Existing Liability Models and Pose New Financial Burdens." *Seattle University Law Review* 42(3). 2019. B2.5.
Tang, Ao, et al. "A Real-Time Hand Posture Recognition System using Deep Neural Networks." *TIST* 6(2). 2015. A2.28.
Thomas, Katie and Charles Ornstein. "Facing Crisis, Sloan Kettering Tells Exec to Hand over Profits from Biotech." *ProPublica*. 2018. B5.2.
Tiana, Vincenzo and Carla Farre Montel. "EU Tech Policy Brief: May 2019 Recap." *CDT*. 2019. B1.8.
Turner, Maurice. "The Threat of Disinformation Looms over EU Parliament Elections." *CDT*. 2019. B1.9.
Ullman, Ellen. *Close to the Machine: Technophilia and its Discontents*. Picador. 2012.
United States. Cong. House. *AI JOBS Act of 2018*. 115th Cong. 2nd sess. HR 4829. B5.1.
United States. Cong. Senate. *Workers' Right to Training Act*. 116th Cong. S 2468.
Vanderelst, Dieter and Alan Winfield. "The Dark Side of Ethical Robots." *arXiv Preprints*. 2016. A1.20.
Wachter, Sandra and Brent Mittelstadt. "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI." *Columbia Business Law Review*. 2019.
Wallach, Wendell, et al. "A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents." *Topics in Cognitive Science* 2. 2010. B4.7.
Wang, Xuewei, et al. "Persuasion for Good: Towards a Personalized Persuasive Dialogue

System for Social Good." *arXiv Preprints*. 2019. A1.22.
Wentworth, James. "The Intelligent Transportation System and the Ethics of Intelligent Computers." *SIGCAS* 33(2). 2003. A2.23.
Williams, Garrath. "Responsibility." *Internet Encyclopedia of Philosophy*. n.d.
van Wynsberghe, Aimee and Scott Robbins. "Critiquing the Reasons for Making Artificial Moral Agents." *Science and Engineering Ethics* 3(25). 2018.
Yamploykiy, Roman and Joshua Fox. "Safety Engineer for Artificial General Intelligence." *Topoi*. 2013. B4.16.
Yershov, A. P. "One View of Man-Machine Interaction." *JACM* 12(3). 1965. A2.3.
Zackova, Eva and Jan Ramportl. "What Might Matter in Autonomous Cars Adoption: First Person versus Third Person Scenarios." *arXiv Preprints*. 2018. A1.23.
.

A Terms and Abbreviations

- AI (Artificial Intelligence): Broadly, the umbrella term for algorithms, machines, etc. artifacts that are considered "intelligent" without direct or continuous human guidance.
- AMA (Artificial Moral Agent): Used in the Machine Ethics literature to refer to machines with ethics built in, I use it to refer more generally (without making claims about what counts as a "moral agent") to AI once embodied in a robot, self-driving car, or etc.
- ICT (Information and Communications Technology): Broadly, the umbrella term for modern communication technologies and platforms, such as telephones, email, social media, etc.
- LAWS (Lethal Autonomous Weapon Systems)
- Machine Learning: AI in the form of fitting insights to data. Contrast this with older Logic Symbolic forms where behavior arises from a pre-determined set of if/then or etc. constructs.

B Methods

Sampling From each database described below, I selected the first ten relevant results or, when fewer than ten were found, all relevant results from the first thirty returned. In a few cases, a result was pay-walled or similar, and I did not go back to find a replacement result. My goal was to reach a convergent concept map of the literature, and the rarity of this pay-wall exception is consistent with that goal.

This initial search produced a total of 131 selected papers.

I first looked on ACM's Digital Library (DL) and arXiv. These are *huge* in computer science. Overlooking them would mean overlooking the opinions of those involved in the technical side of AI. Because these sources slant in general towards papers on how to create AI systems–not papers on "why it matters"–, I broke each down further.

From the ACM DL, I searched under SIGAI (Special Interest Group on AI), SIGCAS (SIG on Computers and Society), JACM (the flagship Journal of ACM), and TIST (Transactions on Intelligent Systems and Technology). Under SIGAI and TIST, I searched for "ethics" or "moral," and under the others I added "AI artificial intelligence" to the query. This produced thirty-three selected papers total from ACM DL.

From arXiv, I searched under Computers and Society (cs.CY), Artificial Intelligence (cs.AI), General Literature (cs.GL), Human-Computer Interaction (cs.HC), and Machine Learning (stat.ML). Each was then searched for "ethics" or "moral" as above. This produced twenty-three selected papers total from arXiv.

For the remaining academic sources, I searched for "AI artificial intelligence," adding "ethics" or "moral" to the query for the first three: Nature (seven results), Phil Papers (ten results), Philosopher's Index (seven results), SSRN (ten results), Hein Online (seven results), and Congressional Research Service (five results).

Two books returned by these searches, but not included in the counts above, are: David Gunkel's 2012 *The Machine Question: Critical Perspectives on AI, Robotics, and Ethics*; and Vincent Muller's 2016 *Fundamental Issues of Artificial Intelligence*.

Finally, I searched as above on three NGOs: EFF (nine results), ProPublica (ten results), and the Center for Democracy and Technology (nine results).

Databases were selected to give a broad overview from different major perspectives.

Reading Papers under forty pages were printed and placed into one of two binders, about 500 pages total in each. The first, "A Binder," contained all papers from ACM DL and arXiv. The second, "B Binder," contained the others. Loosely, "A Binder" represented an account of the views from the technical side of AI, and "B Binder" the sociological side.

I then read "A Binder" starting from the beginning (arXiv) and working towards the end, doing a deep read on each paper, followed by immediately writing a brief annotation and drawing a concept map of the piece. Concept maps were drawn only with respect to the piece in question, and not in a constant comparative way, ie, where the map of an earlier paper was used to give initial structure to a later paper. I read this way until I was confident conceptual convergence had been met, which occurred after the twenty-first paper. The remaining papers in the binder were then given a lighter read with a simpler concept map drawn for each.

For the "B Binder," I used the same strategy as above, but read papers round robin across the databases included. In order to be consistent with the goal of reaching a convergent concept map, I had to address the heterogeneity of the second binder compared to the first: each database in the second binder typically produced pieces from one or two authors, or on one or two central themes, and these themes were not consistent across all databases.

Finally, once after reading the "A Binder" and again for the "B Binder," I drew a large synthesized concept map of the binder's themes and views (stealing one of the classrooms down the hall from my office with the largest chalkboard), with primary organization being dictated by my own synthesized understanding of the literature, ie, not by any one paper's organization.

Writing The large synthesized concept maps were used to organize and guide my writing in the initial "down" draft of this paper. This draft forced me to articulate several points in my own words, connect and synthesize initial ideas, and go back to the source material to check my understanding throughout.

Once the "down" draft was written, I had a better understanding of what this paper was. It was from there, along with discussions with Alan, that I selected the eight views to organize this paper. I used those eight categories to re-structure the entirety of the piece in the "up" draft. Where the "down" draft focused on getting ideas on paper, the "up" draft focused on getting them into order and spelling them out in full, fleshed out detail in the context of that order. This draft took the longest of the three. As I wrote, occasional additional sources were brought in after conversations with Alan, or frequently cited or apparently key sources were further looked into. Longer pieces not printed into the binders were also read at this stage, as I was better prepared to judge whether to keep and how to incorporate those pieces into the paper. That there are eight (a power of two) major divisions coming out of this draft is a coincidence. A ninth (on an epistemological interpretation of trust) was cut to become a future standalone paper.

Finally, in the "surface" draft I took a fine tooth comb to the paper to attend to grammar, flow, clear signposts, table of contents, etc. Drawing from the conventional wisdom of the Composition discipline, make substantial revision past this point, and you are probably writing a different paper. Alan and I then narrowed in over multiple advising meetings on the writing's biggest issues, and a complete and clean re-draft was written after coming up with a thesis-driven outline, included in as an appendix below.

C Limitations

First, papers caught in this review skew towards recent works. This is likely because "artificial intelligence" as a key term is itself relatively recent. Older, more general or analogous, works were only brought in if I was already familiar with them from previous work and the fit was clear or at Alan's recommendation.

Second, the reading and writing process skewed me towards a wide initial search with shallow follow ups. This was in line with my goal of reaching convergence on current views, but it may have also lead me to miss out on a key paper further down the "graph" that could have shed light on one or two issues.

Third, papers skewed towards Western philosophies. This is almost certainly due to searching only in English-dominant databases, made worse by the recency skew if would-be translated works have just not yet come out.

And fourth, discursive formations (the target sense I tried to get in my reading of the literature) are subject to wide-sweeping, sudden, and unpredictable shifts, though otherwise they are generally stable.

The first and second I am confident would not lead to much re-organization of the eight views here and the works discussed under them. The third, however, might. And on the fourth, though these views on AI as a moral matter may look totally alien in fifty years, it

is still a useful backdrop to have at the start of my own study today.

D Outline

1. Introduction
    1. Purpose
    2. Type of review
        1. Focus: Theories (conceptions of AI as a moral matter)
        2. Goal: Identification of Central Issues (the discursive formation on AI as a moral matter)
        3. Perspective: Neutral Representation (resist early ontologization in favor of emerging themes from discursive formation)
        4. Coverage: Exhaustive (til convergence)
        5. Organization: Conceptual (into nine facets of the formation, one excluded here to be attended to in another paper)
        6. Audience: General Scholars (avoiding technical/legal terms as much as possible, worded in lay terms as much as possible, with introductions in each H4, and attempting to convey the sense of each facet of the formation through propositional H2s)
    3. Related canvases
    4. Comparison to/distinction from related canvases
    5. Overview of following sections
        1. Professionalism
        2. Public Trust
        3. Moral Machines
            1. Machine Ethics
            2. Literary (excluded)
        4. Democracy and Societal Effects
        5. Jobs
        6. Consumers
        7. Policy and Guidelines
        8. Personhood
        9. Epistemic Trust (excluded)
2. Body
    1. Professionalism
        1. Theories on Professional Identity
            1. Thesis: Professional identity (non-AI specific) morally matters.
            2. Durkheim (horizontal axis of society)
            3. Burke (identity as rhetorical appeal)
        2. Comparison to Engineering
            1. Thesis: CS/DS/AI as a profession is similar to health and engineering in that it affects the public widely, infrastructurally, and can lead to wide

harms when it fails; it is different in that it is not professionalized.
    2. (non) Professionalization
    3. Software is like a bridge
    4. Ensure public safety by gatekeeping
3. Other professional normative forces
    1. Thesis: Codes of Ethics morally matter because they articulate professional identity.
    2. Examples
    3. Codes of Ethics as articulation of a profession
4. Summary

2. Public Trust
    1. How do we trust AI after a public project has failed?
        1. Thesis: Public AI projects sometimes fail, and the public can lose trust in the AI, the system the AI was embedded in, etc. as a consequence.
        2. Example causes of these failures
        3. Trust-loss might be directed at the system, policymakers, etc. after an AI failure
        4. Example solutions to losing public trust
    2. How do we trust AI in different cultures?
        1. Thesis: Publics must switch to AI from existing non-AI alternatives; to respect their autonomy they must feel comfortable doing this; and what "being comfortable with a moral (software) code" means likely varies by culture.
        2. Public must feel comfortable in order to switch to AI
        3. MIT Moral Machine study results
        4. Opinion that public/AI moral code must be aligned or grokkable
    3. Why is AI different than other technology?
        1. Thesis: AI technology is like existing non-AI technology in many ways as far as assessment and policy is concerned; but it is different in that AI's applicability is vast and its emergent consequences are not well-understood.
        2. What assessments we might consider
        3. Which of those are non-AI specific
        4. Which of those are AI specific
    4. Summary

3. Moral Machines (aka AMAs)
    1. Debate on whether it's morally worth it to make an AMA
        1. Thesis: Having AMAs may make humans less skilled at moral reasoning; but the process of making AMAs may make humans more skilled at moral reasoning.
        2. Opinion that we'll become morally deskilled
        3. Opinion that we'll improve our moral reasoning by making AMAs (TODO add Anderson and Anderson here)
    2. Debate on whether it's possible to make an AMA
        1. Thesis: AI based on data (observations) of human moral behavior and/or

moral preference can be touted as an apparent AMA; it is not an AMA if the AI's design suffers the Naturalistic Fallacy, so the apparent AMA would be a lie to the public that could cause harm if we trust it to make moral decisions; but an AMA's design can still get around the Naturalistic Fallacy even if it is based on data; and there are positive reasons to take the data approach in systems design.
    2. Opinion that AMA research often fails the Naturalist Fallacy check, and so can't be said to be an AMA
    3. Opinion that moral policy derived from data can be evidence of morals (TODO move Data is Tractable stuff here, shorten and clean)
3. Debate on AMA research's focus on true moral dilemmas
    1. Thesis: An AMA should make the right moral choices; in a true moral dilemma there are no good moral choices available; but an AMA is likely not going to face a true moral dilemma; so it is more ethical to design AMAs to seek creative solutions in its time left in the face of an apparent true moral dilemma.
    2. Opinion that it would be better to design AMAs to seek creative solutions to apparent moral dilemmas
    3. ASK ALAN: is there a counter opinion to this?
4. Debate on whether AMAs can make moral decision in best interest for a human
    1. Thesis: AMAs may one day have moral reasoning that appears alien to us or may make moral choices not in any human's best interest; AI can still be used to assist human moral decision making.
    2. Opinion that AMA's dynamical ethics will get out of line with our own (Delacroix)
    3. Opinion that ethical decision making is distributed and therefore machines can't make decisions in best human interest (though they can still assist in human decision making)
5. Debate on whether how we should evaluate AI like or unlike how we evaluate humans
    1. Thesis: When an AI is imitating a human in performing some task and, intuitively, it is hard to evaluate the AI and the human's respective failures the same way, then this is a sign that either (i) we do not understand or need a new way to understand how we evaluate that task or (ii) the AI and the human are not, actually, performing the same task.
    2. Opinion of the basic Turing Test (imitating a human is enough)
    3. Opinion that Turing's "fair play" means a shift in analysis of evaluation of the task in general, not an analysis of whether it's a human or an AI doing it–by default or in most cases
6. Summary

4. Democracy and Societal Effects
    1. AI (in democratic systems) is built on data that is assumed to be representative (aka, data is not neutral)
        1. AI processes vast amounts of human data

1. Thesis: Data matters morally; processing data matters morally; therefore, government institutions cannot use data-based AI as an objective (technocratic) solution to systemic issues of inequality.

2. Data is not neutral
    1. Thesis: Data is not morally neutral.
    2. Human data comes from curation, etc. (Correll)
    3. AI system design makes choices about granularity of data (Garzcarek and Steur; Park and Ghosh)
    4. . . . and how to segment along protected class lines (Kim et al)

3. Data processing is not neutral
    1. Thesis: Data processing is not morally neutral.
    2. . . . and how to measure and aggregate moral preference (Baum)
    3. . . . and how and what objective function to maximize with that data (Overdort et al)
    4. . . . and cannot be equally predictive and fair across groups (Angwin et al 2016b; Larson et al 2016b; CDT 2019)

4. Summary
    1. (rebut) Data-based approaches are tractable. . .
    2. So why is saying "AI is Neutral" bad?

2. AI (in demo. sys.) mediates human discourse
    1. AI-ICTs create new social relationships and/or ways to mediate social relationships
        1. Thesis: Human-human relationships matter morally; ICTs increasingly mediate and create new human-human relationships that matter morally; and AIs increasingly plays a role in such mediating.
    2. AI use in democratic proceedings
        1. Thesis: AI could be used by governments institutions in mediating government-public relationships; this is easier said than done; and failures in these AI systems would likely exclude entire sub-groups.
        2. Big Data government (what it would look like)
        3. Big Data government (issues with that being a reality any time soon)
    3. Misinformation and deep fakes
        1. Thesis: AI could be used with ill intent to sway public opinion and/or disrupt public discourse on matters of high public importance.
    4. Summary

3. AI is being used in specific institutions considered to be of high importance to a democratic society
    1. AI used by non-specialists in high-interest institutions: example, schools
        1. Thesis: The limitations of AI may be hard to understand by non-specialists; non-specialists are increasingly using AI; this includes non-specialists at high-interest institutions.
    2. Why schools have special status (list of laws)
        1. Thesis: Schools are such a high-interest institution.
    3. What that means for AI
        1. Thesis: AI in schools may undermine laws protecting student data

privacy; AI in schools may disrupt the learning environment; and the gains from AI may not warrant such trade-offs.
2. AI is third-party tech with student data
3. AI is invasive
4. Summary

5. Jobs
    1. AI has a wide effect across sectors (list of sectors)
        1. Thesis: AI will replace humans in nearly every sector; this can improve safety, etc.; but this can also displace workers.
    2. Example sectors
        1. Thesis: In health, AI can improve accuracy of diagnoses, but sometimes would be too large a burden on the budget; in national defense, AI can greatly improve intelligence and logistics, but can also be used by political opponents to undermine or blackmail diplomats; and in law, AI can save time and money in locating information, but likely cannot replace humans in tasks like advising and appearing in court.
        2. Health
        3. National Defense
        4. Law
    3. Mitigation strategies
        1. Thesis: The government has policy options to help mitigate the displacement of human workers to AIs.
    4. Summary
6. Consumers
    1. Visible interactions
        1. AI is increasingly playing a part in social situations involving protected class lines
            1. Thesis: AI is increasingly making decisions in situations that involve protected class lines.
        2. AI reproduces structural inequalities because it benefits Whites more than it benefits non-Whites (list of examples)
            1. Thesis: AI may mean an improvement for all regardless of protected class line in those situations; AI likely distributes those improvements unfairly in favor of Whites.
        3. ... because it is not trained on non-White bodies (list of examples)
            1. Thesis: AI may cause harm in those situations where it fails; AI may distribute those failure rates unfairly in favor of those with White body features.
        4. Summary
    2. Invisible interactions
        1. AI is increasingly an "invisible" actor/hand in massive systems
            1. Thesis: Consumers are not aware of the automated actors acting in massive infrastructural systems; this automation obscures the human decision makers in those systems behind a veil of apparent objectivity; (non-) transparency of these automated/human actors is an ethical

decision.
2. AI controlling interfaces
    1. Thesis: AI is increasingly tweaking consumer/business interfaces to maximize objective functions; some see this as automated subtle manipulation of the public in the best interest of the few.
    2. Opinion that because AI can automate maximizing A/B testing it will increasingly manipulate us with little responsibility
    3. Will watchdogs work (if we are like Al Gore's frog)?
3. AI at the system level
    1. Thesis: AI deployed at the system level could lead to illegal behavior emerging from AI-AI interaction and with no clear actor or group of actors to hold accountable.
    2. AI research not tested at system level
    3. AI could automatically collude
4. Summary

7. Policy and Guidelines
    1. Who is making AI policy
        1. Overview (list examples)
            1. Thesis: AI policy-making involves many official actors.
        2. US Government
            1. Thesis: The US government has only recently officially had interest in AI policy.
            2. History (Calo)
            3. Congressional examples
            4. But, government brain drain
        3. AI Standardization (more broadly)
            1. Thesis: AI policy-making broadly investigated involves many official and non-official actors at various scales of complexity and with competing interests.
        4. Summary
    2. What are the goals of AI policy
        1. Stop the bad
            1. Thesis: One goal of AI policy is to limit AI's potential wrongs, harms, and violations of rights.
            2. Prevent harm
            3. Protect privacy
            4. Prevent misuse (COMPAS example)
        2. Promote the good (list example principles)
            1. Thesis: One goal of AI policy is to promote AI and AI systems that have specific, fundamental, and good properties.
        3. Summary
    3. What are some specific AI policy approaches
        1. Limits of Codes of Ethics
            1. Thesis: The AI industry cannot be trusted to regulate itself.
        2. Privacy and GDPR

1. Thesis: To protect and redress consumers with respect to data privacy, AI policy may require attention to high-risk inferences based on data.

3. Patents
    1. Thesis: Overbroad patents for new AI technologies undermine AI policy's innovation goals.
4. LAWS
    1. Thesis: LAWS could revolutionize warfare; humans must retain meaningful control over LAWS.
    2. Meaningful human control
    3. DODD 3000.09
    4. Critique of meaningful human control
5. Tort Law
    1. Thesis: Old technology is not often an appropriate choice of metaphor for regulating AI; one proposal is to use an agency *ex ante* to set standards and certify AI systems and courts *ex post* to handle the complexities of the situation; only system owners with agency certification would face limited liability.
6. Summary

8. Personhood
    1. What is an agent?
        1. Thesis: The definition of a moral agent morally matters, technologically matters, and legally matters.
        2. Standard definition
        3. Fossa's definition
    2. Persons in areas of law
        1. Thesis: An AMA should not be treated the same as a human actor in the cases of criminal, civil, and copyright.
        2. Criminal Liability
        3. Civil Liability
        4. Copyright
    3. What if we give AI moral agent status?
        1. Thesis: It is gross and cruel to murder an AMA; however, that is not ripe, and giving AI moral agent status may lead to a utility monster in no human's best interest that greatly exceeds a fair resource allocation.
        2. Plant
        3. Human Stranger
        4. Murder an AI
        5. Utility Monster
    4. What if we give AI moral patient status?
        1. Thesis: AI, if sufficiently human-like and/or if promising to let humanity live on forever should be given higher moral patiency status than an everyday possession; however, it may be more ethical to not build human-like AI that competes with us for resources.
        2. Bible, flag, Koran
        3. Human Child, or with savant syndrome

        4. Humanity living on forever
        5. Don't make something we'll compete against for resources
    5. Summary
3. Conclusion
    1. Purpose
    2. Type of Review
    3. Connection to related canvases
    4. What this adds to the literature
    5. Distillation of AI as moral matter discursive formation
4. Appendices
    1. Terms and Abbreviations
    2. Methods
        1. Search
        2. Reading
        3. Writing
    3. Limitations
        1. Heavy recency
        2. Wide, not deep
        3. Heavy Western
        4. DFs can suddenly shift
        5. Reflection
    4. Outline

Special Thanks I would like to give thanks to Dr. James Gary for always believing in me and for giving me the resources to study theoretical computer science, and Dr. Mustafa Atici for carrying me up the hill as we learned the parts of my thesis together. I also need to thank the computational artists on my Twitter feed for showing me how everything's done: Janelle Shane, Joanne Hastie, Kjetil Golid, Joanie Leercier, Anders Hoff, and Zach Lieberman. I also want to thank Drs. Peggy Otto, Alexander Poole, Niko Endres, Jane Fife, and Christopher Lewis, who always helped me focus on the good while writing and respectively taught me to approach questions pedagogically, pushed me to study Wittgenstein, or Nietzsche, or Burke, or Morrison. Drs. Reginold Royston, Rebekah Willet, Bianca Baldridge, and Michele Besant I need to thank for always allowing me to bring all of who I am when they lent me an ear. And of course David Price, who printed this draft, and the one before that, and the one before that. As Epicurus declared in his Principle Doctrines and Lucretius defended in *De Rerum Natura*, "nothing enhances our security so much as friendship."