Embracing AI Failures

The potential for AI technologies to enhance human capabilities and improve our lives is of little debate; yet, neither is their potential to cause harm and social disruption. While preventing or minimizing AI biases and harms is justifiably the subject of intense study in academic, industrial and even legal communities, an approach centered on acknowledging and embracing AI-based failures has the potential to shed new light on how to develop and deploy ethical AI-based systems.

Why focus on failures? AI models are our attempts to represent and operationalize key aspects of real world systems. By “attempt” I mean that it is difficult, if not impossible, for AI models to ever fully capture the complexity of the real world. Consequently, errors are essentially inherent to AI models by design. That is, many AI algorithms work to optimize some objective function, such as minimizing some notion of loss or cost. In doing so, and because AI models only partially represent reality, AI algorithms necessarily must trade off errors and sacrifice parts of the input space to produce functions that can generalize to new scenarios. Therefore, while efforts to avoid AI biases and harms are needed, ethical AI development must also recognize failures as inevitable and work towards systematically and proactively identifying, assessing, and mitigating harms that could be caused by such failures.

How should we think about failures? When thinking about AI failures, we need to think holistically about AI-based systems. AI-based systems include AI models (and the datasets and algorithms used to train them), the software and infrastructure supporting the operation of those models within an application, and the application interfaces between those models and the people directly interacting with or indirectly affected by them. AI-based failures therefore go beyond traditional notions of model errors (such as false positives and false negatives) and model accuracy measures (such as precision and recall) and include sociotechnical failures that arise when AI-based systems interact with people in the real world. For example, medical doctors or judges viewing AI-based recommendations to inform decision making may over- or under-estimate the capabilities of the AI components making those recommendations, to the potential detriment of patients and litigants. Acknowledging this type of sociotechnical failure has motivated an exciting and active area of research in algorithm transparency and explanations. Characterizing other types of sociotechnical failures associated with AI-based systems can reveal additional opportunities to mitigate harms.

How can thinking about failures help to mitigate harms? Thinking about potential points of failure in holistic AI-based systems also highlights opportunities for new types of solutions. Consider, for example, developing an AI-based notification system that automatically detects important tasks and sends people reminders when they are due. In this scenario, harms may occur if notifications appear when people are attending to critical tasks like driving. This is not a failure of the AI-model which may be correctly detecting tasks and due dates, it is a failure to adequately consider likely contexts of use. Potential mitigation strategies may therefore include re-architecting the system to monitor, infer and suspend notifications in critical contexts as well as designing mechanisms to support efficient manual dismissal in case notifications still mistakenly fire. Other types of sociotechnical AI failures may include expectation mismatches, careless model updates, and insufficient support for human oversight and control.

Framing AI development around identifying and mitigating sociotechnical AI-based failures may reveal new opportunities to ensure fair and responsible use of AI in society.
Ethical AI Starts with the Data

Michael Zimmer
Marquette University
Milwaukee, WI 53233, USA
michael.zimmer@marquette.edu

Emanuel Moss
CUNY Graduate Center | Data & Society
New York, NY
emanuel@datasociety.net

Position Statement
Scholarship on the ethics of artificial intelligence has been, perhaps justifiably, largely focused on assessing how AI’s outputs might (re)produce instances of bias, unfairness, and injustice. Outputs, whether scientific models, products, or services, are where the ethical consequences of automated decision systems become most apparent. However, numerous potential impacts of AI also stem from the data collection process that precedes any data science or machine learning that takes place subsequently.

While various research communities have recently engaged deeply with the ethics of data collection, sharing and retention practices [1,2,3,4], these debates have often remained less visible to the computer science, data science, and machine learning communities that drive the development of AI, and where a separate set of debates surround the question of AI ethics. This threatens to produce a gap between data ethics and AI ethics with important implications—many of the AI and machine learning applications that have drawn scrutiny for causing unjust outcomes start with data collection practices that fall below ethical thresholds.

Data collection is often where unjust power relations between scientist and research subject are most clearly established, preceding the steps that are deployed to actually develop an AI system. Yet the conditions of
that relationship can be predictive of the ethical stakes of how the science will be applied, either as research or engineering outputs. Three examples illuminate this concern.

**Example 1:** Negative public attention surrounded an algorithm that could infer sexual orientation from photographs and facial recognition algorithms trained on videos of transgender people. In both cases, there were ethical concerns about both the purpose of these algorithms and the fact that the data that trained them (dating profile photos and YouTube videos, respectively) was "public" but collected from potentially vulnerable populations without consent.

**Example 2:** Google’s artificial intelligence company DeepMind has been granted access to the identifiable personal medical information of millions of UK patients through a data-sharing agreement with the Royal Free London NHS Foundation Trust. The agreement provides access to patient data from the last five years, including information about people who are HIV-positive, for instance, as well as details of drug overdoses and abortions. Ethical concerns arise due to the lack of any data anonymization, as well as a lack of specific patient consent for the data sharing. Google has said that each individual patient’s consent for their data being shared is implied, because it is providing "direct care" to Royal Free’s patients through its machine learning tools.

**Example 3:** The NIH National Institute of Mental Health is increasingly funding research to leverage existing electronic health record (EHR) data and advancements in statistical modeling to improve the prediction of suicide attempts over conventional self-reporting methods. Researchers seek to improve existing EHR-based suicide risk prediction models by integrating additional datasets to create multifactorial predictive models of suicidal behavior risk, such as publicly available datasets containing financial, legal, life event and sociodemographic data. Combing such data with medical data represents a potential collapse of longstanding boundaries that shape individuals’ willingness to share data within particular contexts.

Unlike for many other research disciplines, the data collection that fuels the training and testing of machine learning, neural networks, and AI systems generally occurs prior to and separate from the analyses data scientists perform. But that does not absolve them of their responsibilities to ensure that data collection, and data re-use, is utilized in a fair, ethical, and just manner. While data scientists and those building AI systems cannot always control the conditions under which the data they utilize is collected, their use of the data raises a number of challenges and concerns that have not traditionally fallen under the rubric of AI ethics.

Participating in the "Good Systems: Ethical AI for CSCW" workshop will foster conversations about practices of data collection and retention within the emerging practices of pursuing ethical AI.

**Acknowledgements**
This work is supported by the National Science Foundation under grant numbers 1947754 and 1704425, and by the Pervasive Data Ethics for Computational Research (PERVADE) project.
References


To Shape the Future of AI, We Must Understand AI Developers
Karen Boyd, PhD Candidate at the University of Maryland, College Park. klboyd@umd.edu

Researchers have verified algorithmic discrimination in outcomes and accuracy on the basis of age (Diaz et al., 2018), gender (Bolukbasi et al., 2016, Dastin, 2018), race (Sweeney, 2013; Mehrotra et al., 2017; Angwin, 2016), and the intersection of gender and race (Noble, 2018; Buolamwini & Gebru, 2018) across product types, including (for the above examples) text processing, search engines, facial recognition, ad delivery, and criminal risk estimates.

Fairness isn’t the only ethical implication of machine learning (ML): concerns about privacy and accountability have also been raised as well. Anonymized training datasets released publicly have been reidentified (Narayanan & Shmatikov, 2006) and researchers, the popular press, and the courts are discussing algorithmic accountability and due process in big data (Crawford & Schultz, 2014; Angwin, 2016b; Wisconsin v. Loomis, 2016).

In response to ethical concerns, researchers have designed, tested, and published technical and practice-based interventions throughout the ML development process and conferences like ACM’s Fairness Accountability and Transparency (FAT*) and Artificial Intelligence, Ethics, and Society (AIES) have emerged to publish ethics-focused work. But will interventions be adopted?

In my view, the practices, perspectives, and pressures on ML engineers are understudied areas, but are key to making sure that the technologies, policies, and messages concerned actors use to try to influence AI development have their intended effects. I am early in my career and excited to change my mind with data, but here are my impressions of ML engineering based on reading, my own data, and friendly shouting matches with colleagues over beers at ICWSM:

- ML engineering (and software development in general) has an unusually strong occupational culture (Hofstede, 2011), considering its relatively few traditional markings of professionalization (i.e. no consistent training regime, certification, or other formal means of occupational closure)
- Many ML engineers see their job as primarily technical. Even though their inputs, products, and their products’ impact on the world are socio-technical, they (and their educators and employers) scope their job as primarily to achieve technical requirements and improve performance metrics. The jobs of assessing and managing social impact are someone else’s.
- Although social impact isn’t seen as their job, because of their unique skills, understanding, and access to this especially opaque and sophisticated technology, ML engineers may be the only ones in a position to affect the social impact of it by manipulating training data, setting parameters, and communicating to users about the specific risks and weaknesses of a product’s use.
- When ML engineers talk about issues of bias, privacy, and other ethical concerns, they use framings about “quality” and “security.”

I am concerned about increasingly capable AI because of its scale, impact, and opacity. Some users seem to believe that delegating important decisions, like hiring, evaluation, policing, and parole, will get around human bias, when in fact ML entrenches that bias. If we don’t take it seriously and address the social part of this sociotechnical problem throughout development and use, increasingly capable AI systems will cause more harm. My position about the future of AI is that in order to nudge it in a direction that’s aligned with human interests and values, we must understand and consider the lifeworlds of ML engineers.
References


We invited 47 participants and randomly assigned them to three groups that did not receive chatbot’s self-disclosure; received chatbot’s low-level self-disclosure; and received deep self-disclosure, respectively. After using the chatbot for three weeks, we then asked if they were willing to share their self-disclosure data with a real mental health professional through the chatbot. By examining participants’ self-disclosure data one week before and one week after sharing with the real mental health professional, we found that the chatbot’s deep self-disclosure successfully elicited participants’ deeper self-disclosure not only to the chatbot but also to the real mental health professional, more so than other chatbot’s chatting styles. We also found that even though the overall self-disclosure depth was consistent within each group between interacting with the chatbot only and sharing with the professional via chatbot, there were still variances among the details of people’s self-disclosure.

However, several participants felt surprised when they were asked about if their self-disclosure data could be shared with a real doctor. Though these individuals had been introduced to the idea of the doctor’s involvement before being asked to share their answers with the doctor, they still felt surprised. For example: To be honest, I felt offended in the beginning. Maybe when I talked to the chatbot, I thought the conversation was only between the chatbot and me, so I disclosed a lot of secrets. But soon I calmed down and was willing to share my answers because I felt I could trust the doctor. (S35, F)

There are several ethical issues worth addressing when designing a chatbot for health cares, and we want to propose them in this workshop. First, the impact of informing users in advance. When deploying such a chatbot into real applications, it is likely when people start using a chatbot they would know the data might be shared with a professional. This setting may affect users’ self-disclosure behavior with a chatbot because they might feel like they are being monitored and be reluctant to self-disclose deeply, which may deter the efficacy of using a chatbot to improve mental health. Prior studies suggested that people had less concerns of anonymously sharing their stress, depression, and anxiety on online social media platforms, thus anonymity may be leveraged by people when self-disclosing via chatbot to address their concerns shared in the interviews.

Second, should the AI-agent be designed to dig for users’ private information? CASA paradigm indicated that people may mindlessly apply human social norms to a computer agent. Thus, people might disclose some highly intimate information to an agent after gaining trust in the agent; however, if disclosed content includes criminal confessions and unreported harassment issues, it will be a dilemma for a chatbot system to deal with this situation. A chatbot’s user might be informed that their conversation is confidential, so the user may be willing to disclose more. But if the chatbot detects something illegal, should the chatbot brake the promise? There is a lack of legislation to address this conflict, but if it is possible to disclose information without permission, the users would be concerned and disclose less information. In addition, validation of the disclosed content could also be a concern for reporting the issues. In brief, these ethical issues need further discussion in the workshop.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI’20, April 25–30, 2020, Honolulu, HI, USA
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-6708-0/20/04. . . $15.00
DOI: https://doi.org/10.1145/3313831.XXXXXX
Consider the following about the popular crowdsourcing website Amazon Mechanical Turk. Though the site is widely used in the world of research, the question of who is doing the click work is quite often overlooked, with the assumption that everyone on it is of legal age to work because it is required of Turkers to be 18. However, there is no verification process to ensure that the workers are in fact the age they claim. Go a little ways down the Reddit hole and you will quickly find threads of users admitting to, and speaking to the ease at which they are making money on MTurk while being underage. Such breaches of labor laws (along with the evidence) seem to go completely ignored because, afterall, what can really be done? The Internet knows though – should it be doing something? In the future it does.

In 1986, Langdon Winner addressed the challenge of having to search for limits in an age of high technology in The Whale and the Reactor [2]. That was over 30 years ago and the technological peak we are positioned at no longer mandates us to keep searching aimlessly for the limits. We can and are on the brink of hard-coding society’s moral boundaries and digitizing ethical standards. Look to the Moral Machine experiment as an example of a crowdsourced platform for gathering a human perspective on moral decisions that machines will eventually be in charge of making, such as self-driving cars [1]. If most of the world thinks that a car ought to swerve to kill a very old person to save a pregnant woman’s life, then it seems logical that the world’s self-driving cars will remain consistent with that judgement. Following suit, we can collect, quantify, and integrate the societal expectations around different harmful activities we observe in our systems. For example, how ought a teenager who sits down at their computer every day to cyberbully strangers on the Internet be dealt with? We have the ability to design systems that are capable of condemning such activities if we collectively agree that they break some moral standards.

A sentient web sits at the intersection of artificial, collective, and emotional intelligence. It is a world of digitized support systems and computers that can condemn. Different situations garner different moral judgements. Once we formalize all aspects of our own emotional intelligence, our conceptions around empathy in their entirety, we allow for sentient systems that are guided by a collective moral compass to moderate online interactions, and to ensure our technologies serve as positive “forms of life” that make way for building order in the world. [2]
References

The future I’d like to explore is one in which AI Systems are evaluated as ethical agents, rather than as mere tools employed by human ethical agents. This will require two main conceptual innovations:

- **An account of algorithmic agency.** How can AI systems be said to do something, and how can they be held responsible for these actions? Do the actions of AI systems have ethical relevance for others, such as those that build, design, or employ them?
- **An account of (philosophical) ethics suitable for AI systems.** This must go beyond the familiar code-of-ethics based folkways. While folkways and community norms can be important forces for good, they are often developed in reaction to undesired behavior, which diminishes their ability to consider new technologies and practices. A coherent philosophical way of evaluating whether AI systems’ actions are good is needed to proactively and prescriptively circumscribe and guide their development.

My own work is addressing these two questions in ways I hope prove generative.

My dissertation attempts to clarify the actions and actors that enable information systems to condition social reality. Databases, for instance, now routinely contain facts, in the sense of propositional logic. That is, instead of representing some state of the world, databases such as airline ticketing systems, the No Fly list, and China’s Social Credit Score directly create the states which determine whether statements are true or false. This shift is not unique to computerized information systems; written records have sometimes played analogous roles before. But the proliferation of these systems has made this phenomenon more widespread, and more portentous in recent years.

Key to this work has been identifying the actions of *promulgation*, the making of inscriptions within information systems, and *enactment*, the conditioning of action based on promulgations. Each process has a corresponding agent. Promulgating agents are increasingly automated systems: China’s Social Credit Score is computed by algorithm, for instance. Automated enacting agents are becoming more prevalent as well. Whereas a conversation with a bank loan officer may have once been required to get a loan, and in China such officers would thereby become enacting agents that translated the Social Credit Score into social reality, it is common that even the enactment of these systems is becoming automated as well. This provides an end-to-end platform whereby those with control over automated promulgating and enacting agents can condition social reality. Human agents retain a role in many of these systems, and form a complex ethical assemblage that I’ll address in ongoing work.

My ethical approach to this problem has been through the work of philosopher Richard Rorty.¹ Rorty attempted to build a pragmatic account of ethics that was coherent even when confronted with the radical historical and cultural contingency of ethics as a philosophical project. In an increasingly divided and at times “post-truth” global context, Rorty’s thought offers an inroads to the problem of the ethical character of AI systems, their actions, and their manifold imbrications with human actors. He advances the principle of Solidarity as a way of reclaiming and justifying a focus on the value of human life, even for those that realize the contingency of ethics.

Regardless of the success of my approaches, I hope to build consensus around the importance of the questions. Considering them thoughtfully will help us avoid playing ethical catch up to AI systems.

---

Planning for Ethical Agent-Agent Interaction

Jesse David Dinneen
Victoria University of Wellington, New Zealand
jesse.dinneen@vuw.ac.nz

ABSTRACT
In this position paper for the 2019 CSCW workshop Good Systems: Ethical AI for CSCW I propose one tool and one idea for navigating the complex ethical problem space that results from the interaction of human and/or AI agents in shared, hopefully cooperative, computing environments.

KEYWORDS
AI ethics; human-computer interaction; ontic trust

AGENT-AGENT INTERACTION
The introduction of non-human agents (e.g., AI-powered virtual assistants like Apple's Siri), which are increasingly indistinguishable from human agents, to everyday computing raises many questions about what AI, our interactions with it, and AI-AI interactions (all collectively: agent-agent interactions) could and should be like, especially as AI grow in capacity to be moral agents akin to humans (Floridi & Sanders, 2004). To guide implementations and expectations of moral AI requires considering the many ways agents’ actions can be undesirable, but presently most news media and even scholarly discourses focus narrowly on poor performance, transgressions of the law, and negative outcomes for particular individuals irrespective of their socio-economic group or status (Stahl et al., 2016). Thus, while the moral-problem space of agent-agent interaction is arguably more vast than that of human-human interaction, we designers and critics have so far used fewer tools to analyse it.

Towards addressing narrow thinking in AI ethics, and to aid anticipatory (rather than reactionary) policy, I would like to consider the use of a multi-moral matrix for assessing (in design or post hoc) particular AI cases along multiple moral frameworks. Figure 1 shows an example matrix with agent actions in the leftmost column and each other column showing possible transgressions according to various moral frameworks (summarised to the point of caricature), including (left to right): legalism, consequentialism, virtue ethics and deontology, social justice, and social contractualism. Cells at the intersections of actions and frameworks may reflect only that there are possible transgressions or may contain more detail. Agent actions may be relatively easy to populate, e.g., by identifying actions in user stories during development.

The example matrix includes only a few moral frameworks and is meant to be neither exhaustive nor prescriptive; any use of such a matrix requires customisation according to the expectations of, e.g., the relevant sectors and cultures.

Finally, I suggest to adopt or adapt into thinking about AI ethics an idea that appears thematically appropriate and also promising for addressing cultural moral differences like those just mentioned (Hongladarom, 2008): the concept of ontic trust (Floridi, 2009). Put briefly, ontic trust is a responsibility to care for the intrinsically valuable information objects that populate our world/infosphere; shorter still, causing entropy in a shared information environment is unethical. Such an idea shows us a less obvious way agent-agent interactions can be bad (i.e., by causing entropy – it could thus go in the above matrix), but arguably also implies rights for non-human agents (i.e., AI).

Surprisingly, little has been said about ontic trust in AI ethics discourse (and to my knowledge, nothing has been said in the context of CSCW). It may therefore be useful to raise the idea at the workshop and discuss questions like:

- Is an imperative to care for information objects equivalent to an imperative to prevent entropy?
- What do such imperatives imply entail for privacy, data logging, the right to be forgotten, and CSCW community values? (Bruckman et al., 2017)
- Can ontic trust aid universal design by, for example, mediating disparate cultural views about ethical AI?

Figure 1. Partial example of a multi-moral matrix for analysing ethical issues in agent-agent interaction.
REFERENCES


Ethical Concerns for the Future of Face Recognition Technology and Policy
Kolina Koltai (UT Austin), Claudia Flores-Saviaga (West Virginia University), Megan Rim (University of Michigan)

Regulation of face recognition technology (FRT) has not kept pace with increasingly widespread adoption of face recognition technologies. This has resulted in the unregulated use of FRT by law enforcement in places such as Detroit, Michigan, and Baltimore, Maryland. While several bills have been introduced, there currently exists no comprehensive protections at the federal level against the use of face recognition on citizens. Although, San Francisco, California, Somerville, Massachusetts, and Oakland, California have instituted bans on FRT, some cities have opted instead to pass regulatory policies face recognition to the effect of imposing some restrictions but ultimately legitimizing its use as a policing technology. As some cities have begun implementing real-time face recognition technologies, this has stirred debate about how its use by law enforcement might violate the First Amendment’s protections of the right to freedom of assembly, and Fourth Amendment’s protections against the unlawful search of private spaces (Hamann and Smith, 2019).

As it stands, we do not anticipate FRT to leave the public space in the near future. We anticipate that more governments and agencies would want to incorporate FRT into monitoring, policing, and military efforts. FRT will be built into the infrastructure of society, physically with cameras, institutionally with policy, and culturally as an accepted norm. We also anticipate there being a lack of separation between FRT data and personal data from other companies and organizations. You can imagine Social Media Platform data and FRT data combining to give a ‘catered’ experience when you go to different physical locations. While it may seem ‘Minority Report’-esque, we may not be far from a future that has highly integrated FRT into society from shopping and catered ad preferences and consistent monitoring.

FRT biometrically identifies you by matching your unique facial dimensions against huge databases. However, a recent study uncovered large gender and racial bias in commercial Facial recognition software. In the researchers’ experiments, the accuracy in determining the gender of light-skinned men were never worse than 0.8 percent. For darker-skinned women, however, the error rates ballooned to 35 percent (Buolamwini, 2018). Afterall, FRT is only as smart as the data used to train it. If the system is trained using faces of many more white men than people of color, then it will be worse at identifying these minorities. This is worrisome as across the U.S, state and local police departments are building their own face recognition systems. But, we know very little about them. i.e. we don’t know how they address accuracy problems. As a consequence, we don’t know how any of these systems affect racial and ethnic minorities.

Recent research has proposed ways to reduce bias in identifying people in different demographic groups (Amini, 2019), but without regulation, that won’t curb the technology’s potential for abuse. Ultimately, as accuracy is improved and bias is mitigated, it is expected that law enforcement will want to use FRT for immediate identification. For example, it might soon be possible to scan the faces of people passing by the street using CCTV cameras and determine not just who someone is, but where they’ve been, where they’re going, and whether they have an outstanding warrant, immigration detainer, or unpaid traffic ticket (Kofman, 2017). If FT systems that government and law enforcement agencies use is biased and with low accuracy, there is risk that the face recognition search will lead to an investigation, if not an arrest, of the wrong person (Garvie, 2019).

The main problem is that existing privacy and civil rights laws were mainly designed to limit old-fashioned forms of privacy violation, such as illegal searches or unauthorized revelation of private activities, such as medical records. Currently, there is evidence about how face recognition is being used in police surveillance of protests (Garvie, 2016). This could have an impact on the public and political discourse. For example, past research has found that surveillance practices may create a chilling effect on democratic discourse by stifling the expression of minority political views (Stoycheff, 2016). If the of face recognition technology in public spaces continues to expand, minorities might not choose to participate in activities such as protests, if they know their face could be scanned. In the absence of regulation, the use of face recognition for law enforcement, could lead to serious risks of misidentification. In the absence of transparency, these uses threaten to violate the due process rights of those arrested (Garvie, 2019).
References


What’s in a face?
Speculating the future of Computer Vision through the lens of AI

Shivani Kapania & Anushka Bhandari *
shivanik@iiitd.ac.in, anushka16134@iiitd.ac.in
IIIT Delhi, India

As researchers who have worked both in Computer Vision (CV) and HCI, we recognize that CV applications are becoming ubiquitous day-by-day due to the vast (and ongoing) research in machine learning techniques, especially, convolutional neural networks [1] and generative adversarial networks [2]. The computer vision community has developed methods for object detection and tracking [3], facial recognition [4], gesture recognition [5], and much more. These systems are being implemented by small scale companies/ startups and multi-national corporations like Amazon, Google, and Facebook alike [6]. However, with the rapidly growing influence of all such applications, as researchers and practitioners, we must be wary of all the impending ramifications that may arise in the near or distant future.

In this paper, we draw on the work of Skirpan and Yeh [7] by investigating the Present and speculating the future of Computer Vision. Time and again, we have been informed about the socio-technical implications of AI systems: the invasion of users’ privacy, the possibility of security breaches, and most commonly, discrimination against individuals belonging to under-represented groups [8]. In 2018, Buolamwini and Gebru uncovered intrinsic racial and gender bias in facial recognition services by IBM, Microsoft, and Amazon [9]. More specifically, while these systems had high overall accuracy; when the researchers analyzed the results by intersectional subgroups of skin type and gender, these systems performed worst on darker females. Furthermore, CV methods capable of tracking minority groups are being used for surveillance (guiding police response) which undermines their privacy and raises questions about discrimination against these groups [10].

By attempting to locate and address the above gender or racial biases individually, we overlook the social constructs (who creates? who benefits? whom it harms and how?) which determined these results in the first place. While we advocate for human-centered AI approaches, we also acknowledge that policy-level reforms are required to bring about fairness, accountability, and transparency in these systems. As Eubanks pointed out, "When automated decision-making tools are not built to explicitly dismantle structural inequities, their speed and scale intensify them.” [11].

To this end, we suggest a three-pronged approach towards more ethical AI:

- **Inclusive AI workplaces:** AI reflects the values of its creators. Research has pointed out that there exists a feedback loop between the (discriminatory) workplaces and the AI tools they build [12]. Tech giants like Google and Facebook have only 10% and 18% of women researchers [13]. By addressing the diversity problem in the AI workforce not just for gender and race, but also power inequities, we can create systems better suited to operate in high stakes scenarios.

- **Questioning data and algorithmic biases:** Data lies at the heart of AI systems. This historical data may be incomplete and non-representative of the current social structures. In 2018, Uber suspended the accounts of transgender drivers over a security concern raised by its facial recognition system [14]. Most of these systems assume a gender binary and have difficulty identifying individuals undergoing gender transitions. Researchers and practitioners must en-

*Both authors have contributed equally.*
sure that data collection practices respect the varied demographics and contexts of its users, and address the finer yet invisible issues from diverse vantage points.

- **Policy-level reforms:** By accepting that the definition of ethics is relative, continually evolving, and repeatedly challenged, we understand that developing a "perfectly ethical" AI seems implausible. In contrast, by placing checks using concrete policy frameworks on data-driven decision making and in general, AI, we can ensure that power asymmetries don’t reinforce and perpetuate inequality through these systems.

We acknowledge that our interpretation of this line of work is shaped by our educational backgrounds, programming experiences, and personal perspectives. All the authors are of Indian origin and are exploring approaches for designing systems for marginalized groups. One author has examined the design of more inclusive mobile applications for low-literate users. One author has worked on developing a learning platform for community health workers. Both authors have experience developing deep learning AI models for multiple object tracking and super resolution in images. Attending this workshop will allow us to gauge the opportunities and challenges in designing and developing ethical AI systems and participate in discourses related to our areas of research.

References


Ethical Considerations for Adolescent Online Risk Detection AI Systems

Afsaneh Razi  
University of Central Florida, USA  
afsaneh.razi@knight.ucf.edu

Seunghyun Kim  
Georgia Tech, USA  
skim888@gatech.edu

Munmun De Choudhury  
Georgia Tech, USA  
munmund@gatech.edu

Pamela Wisniewski  
University of Central Florida, USA  
pamwis@ucf.edu

ABSTRACT

We seek to develop future Artificial Intelligent (AI) risk detection algorithms to address keeping adolescents safe by providing accurate and customized services to teens and their parents. Training such accurate algorithms needs data of minor which raise ethical challenges. Also, the proper use of such systems is another issue. In this workshop, we hope to gain more insights about possible approaches to address these ethical challenges.

Author Keywords

Ethical AI; Online risk detection; Adolescent Online Safety.

CSS Concepts

• Human-centered computing–Human computer interaction (HCI)

GRAND ETHICAL CHALLENGES FOR AI ADOLESCENT ONLINE SAFETY

Internet and social media use is increasingly deeply intertwined in teens’ lives [11]. Although it provides a great opportunity for teens to learn, it also exposes teens to various online risks [6]. Research has shown that solutions for adolescent online safety are relying more on parental control through device-based restrictions and direct monitoring [3]. The current approaches overwhelm parents with teens’ information which they do not find it useful, and also are privacy-invasive to teens [3, 4]. The reason is current risk detection algorithms do not take the context of online interactions into consideration (e.g. teens using curse words when joking in a group chat is different than using such words directly targeted to someone). Thus, machine learning risk detection algorithms should be optimized to the actual content that parents and teens find it risky. In the future, these optimized systems with the use of AI will help adolescent and their parent have a safe online experience. We have an NSF funded project [7] to improve automated risk detection algorithms using human-centered principles. We plan to commercialize the final solution for an easy-to-use and accessible service for adolescent online risk detection.

Using Teen Data to Create ML Algorithms

In our project, we are collecting social media data of teens which includes both their private and public data. Since training with representative data is a big part of making effective classifiers, we need a contextual and in-depth knowledge of adolescent risk behaviors [9, 10]. Current risk detection systems are not taking into consideration the nuance in risk classification when it comes to risky social media content of teens [1, 5]. So, we are using human-centered approaches and qualitative analysis to develop risks concepts. We aim to label teens social media data for ground truth before developing advanced and automated algorithms for risk detection.

However, the data collected from teens might include sensitive and possibly illegal artifacts, such as sexually explicit images that could be classified as child pornography [8]. These sensitive data pose serious ethical issues for the human-centered approach to building AI risk detection systems. Thus, we need to ensure that our data collection and algorithmic analysis of teen social media data is ethical. We have considered some ways to address these ethical issues. In collecting teens data, not only we obtain parental consent, but also, we get teen assent. We believe it gives teens more sense of authority and control over their data. There is also an intrinsic challenge in the process of collecting the data, since asking the teens to self-identify harassments might result in a recap of those experiences. We make sure to not inflict a time constraint to the participants when collecting data to not possibly overwhelm them with such experiences. For preserving the privacy of the users, we make sure to not publish any of the personally identifiable information or any quotes of their messages that can be retrieved through online search. Also, we convert usernames to randomly generated IDs to protect the privacy of the users.

Using teen data to create ML algorithms.

Deploying Adolescent Online Risk Detection Algorithms for Good (Not Evil)

After the AI system is built, it is important to protect it from the wrong hands. For instance, if there exists a system that can detect minors sexually explicit images, some people can misuse the system and use it as a tool to find sexual images of youth to save it in a porn website or in darknet. Also, the data detected as risky by these systems can be hacked or misused. To combat against the potential reverse engineering of our trained adolescent online risk algorithms [2], the trained algorithms will not be released as open-source. Access privileges to the system should be monitored in order to ensure that only the correct people have the right type of access to the system. Thus, safeguards should be devised to protect the security and privacy of such system and more research should address these challenges.
CONCLUSION
We found approaches to address some of the ethical challenges for improving adolescent risk detection systems. As we want to make sure to maintain the confidentiality and privacy of the data and security of the AI system that we design. We hope to gain more insights from participating in the Good Systems: Ethical AI for CSCW to tackle challenges for designing AI systems that can promote online safety for teens.

ACKNOWLEDGMENTS
This research is supported by the U.S. National Science Foundation under grant number #IIP-1827700. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors.

REFERENCES
Online health communities and AI: Promise and concerns

Yung-Sheng Chang  
School of Information  
The University of Texas at Austin  
ysechang@utexas.edu

Kolina Koltaï  
School of Information  
The University of Texas at Austin  
koltai@utexas.edu

After the introduction of Web 2.0, online health communities (OHC), such as social networking sites, online forums, and group chats, have increasingly become an information source for health consumers to seek and provide information [1]. The motivations for health consumers to visit OHC include seeking social support, first-hand experiences shared by similar others, and providing information out of empathy and altruism [2]. Nevertheless, OHC mostly contains information provided by laypeople. Users may also provide information anonymously, causing a decrease in identity transparency [3]. Thus, the information in OHC is mostly evaluated as lacking objectivity, trustworthiness, and expertise [4]. With the rapid development and implementation of artificial intelligence (AI) in cooperative work environments, we believe AI can help solve some of the existing problems in OHC as AI has already been shown to improve online health communication [5].

AI has the potential to improve information dissemination on OHC. For instance, AI can provide personalized information to health consumers based on their different information needs, goals, and eHealth literacy levels. Since OHC are mostly laypeople participating in the online conversation, it is likely that health consumers may accidentally provide misinformation. Thus, AI may be able to identify misinformation and provide the correct information instead. For some OHC, especially online forums, health consumers can provide information anonymously. However, anonymity is a double-edged sword that can facilitate information providers’ self-disclosure but hinder information seekers’ credibility judgment due to the lack of identity transparency. With AI, it has the potential to leverage the conundrum by allowing information providers to stay anonymous while providing additional information to verify their identity. Lastly, in OHC set up by private companies (e.g. Facebook), AI can affect the availability of online groups. For instance, if a private company wanted to limit OHC for trans-people looking for health advice and information, they could; AI could help find and identify those groups. Private companies are still free to regulate what content is on their platforms and who has access to it. For example, hospitals with religious affiliations have withheld care from patients because it goes against their beliefs, like when abortions are withheld even in life-saving situations.

Despite the beneficial promises that AI may provide, AI can also bring ethical concerns. AI may violate health consumers’ privacy and data rights. For example, there are no regulations on where the health data shared on OHC are stored or who has access. Companies developing AI program are not gaining consent from participants in OHC for their data to be scraped or analyzed to be used with an AI program. Some OHC may have health professionals to help moderate the information. If AI were to be implemented over humans, it could possibly affect health professionals’ autonomy, reputation, and professional status in the community. Companies might try to mine health data from these OHC and use the data against health consumers’ will. On the macro level, health consumers are now having different accesses to digital technologies due to their socioeconomic status and individual characteristics. Since AI relies on online technologies, it may exacerbate the digital divide, leading to an increase in health disparity. AI is an inevitability in OHC, but it will be critical to be conscientious of the ethical concerns that come with it.
Reference
Good Systems: Ethical AI for CSCW -- Position Paper

Izuki Matuba  
izuki@integrate.ai  
Integrate.ai

Divya Sivasankaran  
divya@integrate.ai  
Integrate.ai

It is indispensable for future AI to have an ability to audit and fix unfairness while securing privacy of users. In the exploration of fixing bias and unfairness in the result of prediction, we found that it is nearly impossible to evaluate the fairness of prediction without exposing each user’s privacy sensitive information such as the one listed above, on which unfair discrimination aligns with. While privacy and prediction performance are usually juxtaposed together as the two factors to balance, our view is that it is an evaluation of fairness which should be considered to stand together with the privacy issue.

Any machine learning model we build is inherently learning to segment the data. Sometimes the data is about equipment in a factory, but in many cases the data is human. That means the point of the model is to divide individuals or populations by some set of attributes. Some of the attributes will be personal or sensitive information, which will rightly trigger privacy concerns. Furthermore, classifying people by these attributes—or others—could cause could cause biased results and lead to unfair impact due to the action taken based on the result.

Attributes are not equal and not all bias is problematic: history and social context matter. The most important fairness issues arise most when segmentation results in negative consequences for members of populations that have long faced systemic disadvantages. Many of these are recorded as protected attributes, for example, gender, socio-economic status, religion, race, and ability status.

Here, we would like to bring an example as a company applying AI technologies to business intelligence domain. One Telco company is trying to actively reach out the customers who are likely to churn from their subscription to provide some support or beneficial treatment. In order to predict which customers are in the risk of cancelling the service, the company consults a third party who is specialized in applying machine learning models to these tasks. However, the Telco company has restrictions on the data to provide as features of the model due to its privacy policy. The third party can still build a model using non person identifiable information, but no longer has an ability to evaluate the fairness of the result despite the fact they are willing to.

We are building a platform to solve the problem in this situation. In some occasions where personal information is not allowed to be used, the platform can still audit the fairness of the prediction result and action taken based on the prediction. Although our AI community is still in a very early stage of this attempt, we strongly believe that the future AI must embody this nature.
Fact-checking: Role of Transparency in Information Access

Anubrata Das  
University of Texas at Austin  
anubrata@utexas.edu

Soumyajit Gupta  
University of Texas at Austin  
smjtgupta@utexas.edu

Alex Braylan  
University of Texas at Austin  
braylan@cs.utexas.edu

ABSTRACT
Misinformation is a growing area of concern for technologists, journalists, policymakers and the government. In this article we discuss how ethical challenges in AI systems, specifically transparency, play a crucial role in battling misinformation by designing and developing fact-checking systems. We seek to answer the broader questions we found in some of our current work.

Keywords  Fact-checking, Information Retrieval, Transparency

1 INTRODUCTION
Information Retrieval (IR) systems play a crucial role as a medium of information access. For a majority of users, these search engines are the primary information system to affirm the veracity of information (Fact Checking). The impact of search engines extend beyond the digital world and can influence the social and political choices of the public. For example, Epstein and Robertson [7] show that manipulated search engine results can shift voting opinion of up to 20% of users.

While interacting with search engines, users have no visibility into the mechanisms governing the ranking of results or the tailoring to the specific user’s search patterns. Transparency in IR can empower users to make better decisions. In this article, we explore the transparency aspect of search systems and how it is relevant in the context of fact-checking.

Research Questions. The complex issue of misinformation requires a multidisciplinary approach. However, information access is at the core of the problem and interests all disciplines such an information, communication and computer science. We are primarily interested in the questions: RQ1 How transparency of Information Access systems affect users’ trustworthiness for a fact-checking scenario? RQ2 How are Information Access systems used for fact-checking?

2 RELATED WORKS
IR systems are one of the most used form of information access systems in terms of seeking information online and have direct influence on user opinion formation.

Political Influence. Studies show that candidate selection for a political scenario can be influenced by search engines [6], and that simple manipulation of search results can change voting choices in undecided voters [7]. Employing a ‘Wizard of Oz’ experiment, Epstein and Robertson [7] show that through communication of existing biases in the system, it is possible to lower the influence of these biases on the users’ decision-making process. Additionally, biases in social media systems can emerge from both the data used to training and design of algorithms [8]. In the context of the presidential primaries in 2016, a study [8] show that 70% of the time top contents in systems like Twitter and Google matches the user’s own political biases. Using a source-based bias-quantification framework led to balanced and less biased results in such systems.

3 DISCUSSION
Fact-checking using computational techniques has the potential to incorporate and amplify existing biases in society. Such systems can, in turn, manipulate users’ socio-political decision making. IR systems can establish credibility in fact-checking when they are fair, free of biases, and interpretable by non-expert users. Looking back to our research questions, we can say that intelligent systems such as search engines can potentially provide a toolbox for fact-checking when they are unbiased and transparent. We see a common theme of methodology: interpretable machine learning. We argue that instead of high-accuracy block models, interpretable models and visualizations are the keys to building reliable decision support systems in the context of fact-checking.

The directions we envision are three fold: 1) Developing Transparent IR algorithms for fact checking. 2) Designing interfaces that enable an unbiased representation of retrieved content. 3) Evaluating the effectiveness of these algorithms and interfaces. Joint expertise from the field of HCI, IR and Journalism are required in order to address these issues.

We conclude by mentioning some of our ongoing research which focuses on two aspects. First, we develop a user interface to communicate the effect of user bias while the user is engaged in a claim checking activity [3–5, 12]. We show that communicating user bias enables the user in making a better judgment toward the veracity of a claim. Second, we develop a framework to evaluate the presence of ideological bias of a ranking system. Given a known target distribution of political ideologies, we estimate a ranking distribution and identify the difference between the estimated ranking distribution and target distribution [2]. Finally, as future work, we aim to develop human-in-the-loop evaluation techniques for measuring transparency and effectiveness of IR systems in fact-checking.
REFERENCES


Communication Design to Navigate the Future of Work and Artificial Intelligence

Joshua B. Barbour, Shelbey L. Rolison, Jared T. Jensen; University of Texas at Austin; barbourjosh@utexas.edu

Forecasting the future of advancements in artificial intelligence (AI) has captured the public sphere particularly as it relates to automation and the future of work (Brynjolfsson & McAfee, 2016; Lepore, 2019; Mukherjee, 2017). The mix of hype and anxiety will be familiar to scholars of the history of technology, but predicting what will, could, or should be the future of work and AI is nonetheless an important endeavor (Stone et al., 2016). However, because of the difficulty and fallibility of such prognostications, attention should also be given to societal resources for navigating the opportunities and challenges of AI and automation as they unfold. Because so much of this technological change will be negotiated by workers, society must also understand how to empower the effective design of organizational communication processes for deliberating about the development and implementation of AI and automation.

The focus on work and workers is important too because new technologies of work can transform it in ways that obscure the transformation as it happens (Leonardi, 2012; Sennett, 2009). The process of automating work also generates information about the work being automated, and the ongoing operation of data-intensive automation also generates a flow of information about the work being done (Zuboff, 1988). The datafication of work and AI may heightens the reach and scope of its ongoing transformation (Brynjolfsson & McAfee, 2016). At its core though, automation transforms work, because changes in the technologies and organization of work are intertwined (Bailey & Leonardi, 2015); and these changes unfold in and through organizational communication (Barbour, Treem, & Kolar, 2018; Leonardi, 2012)

Key questions center on how to shape the communication through which the automation of work occurs and how to cultivate preferred forms of communication. These are questions of collective communication design (Barbour, Gill, & Barge, 2018), or efforts to grapple with the puzzle of “how to make communication possible that was once difficult, impossible or unimagined” (Aakhus, 2007, p. 112). Important communicative phenomena in the intertwined transformation of technology and work include (a) information seeking about new the technologies, (b) “benchmarking” or co-workers’ questions about technology and how to use it, and (c) “technical teaching” or interactions with designers and technologists especially related to how technologies work in practice, (d) rhetorical framing and pitching of new technologies, and (e) policy-focused deliberations in the decision making about what technologies to implement and how to do so (Leonardi, 2012).

The Automation Policy and Research Organizing Network (APRON, https://www.apronlab.org) aims to advance the communicative study of the future of work by researching (a) how technology, organizations, and work change together and (b) the datafication and automation of work. The current research of the APRON Lab includes multiple empirical projects in the context of health and healthcare work. We look forward to the workshop as an opportunity to inform and improve this scholarship.

Funding Acknowledgement
This material is based upon work supported by the National Science Foundation under Grant No. SES-1750731.
References


AI Ethics, Autonomous AI Machines and Global Capitalist Society
CSCW 2019
Position Paper

Ramon Diab, PhD Candidate
Western University
Faculty of Information and Media Studies (FIMS)
Library and Information Science Program
London, Ontario, Canada
rdiab4@uwo.ca
The technical capabilities of advanced artificial intelligence (AI) has generated approaches to AI ethics that consider the implications of the elevation of machine actions and decisions over human actions and decisions. Current AI ethical considerations tend to focus on the potential of AI machines to take control of society, to enact human biases, to enact operations that such machines were not programmed to operationalize, or to enact operations as programmed with unintended consequences. AI ethics tend to focus on the development and embodiment of ethical and legal standards in AI machines during the design phase of AI development and/or the analysis of the consequences of either human-controlled AI or autonomous AI machines. However, the broader historical development of AI ethics has emerged with the historical development of the social forms, relations, institutions and the legal and political superstructure of the state. Current AI ethics therefore abstracts from a critique of the private ownership and development of AI machines within the relation of labour to owners of the means of production and the representatives of the state. The resolution of ethical dilemmas and the secure and safely-controlled development of AI is therefore necessary but insufficient for a broader historical analysis of the implications of the development of AI machines within the global capitalist economy. Thus the use of AI as an instrument of capital, bourgeois institutions and the legal and political superstructure of the state should become an object of critique.

The production and implementation of AI machines within the relations of global capitalist society holds significant implications for the future of work and class. With the historical development of autonomous AI machines at the point of production, the reduction or elimination of the relation of capital to labour-power, and the reduction or elimination of the relation of labour-power to consumer leads to the autonomization of the capitalist mode of production, and thus, to the development of the direct relation of consumers to autonomous AI machines. In the short-term, the development of autonomous commodity production and circulation contains implications for unemployment, wealth inequality, the global reproduction of class, and subsequent social and political responses to the uneven nature of global AI-powered capitalist development. In the long-term, the total replacement of labour-power with machine-power holds significant implications for the relation of labour to capital that is the foundational relation of the capitalist mode of production. Further, the advanced development and implementation of AI machines within the relations of the political and legal superstructure of the state holds implications for the expanded reproduction of the private ownership of the means of production, the private prison-industrial system and the means of warfare. Alternatively, a critique of the bourgeois form of AI development could advance concepts of common ownership and the socialized development of AI as a means of advancing an autonomous mode of production that leads to the abolition of labour-power, the equal distribution of the social product, the dissolution of class and the dissolution of the state.

---

THE ETHICAL IMPLICATIONS OF THE USE OF MATHEMATICAL MODELS AND
ALGORITHMS IN CREW SCHEDULING FOR LONG HAUL TRAVEL

Lore Benson, MSIS Candidate

University of Texas at Austin
Background

In the growing industry of long haul travel, companies are expanding their crews to gather more market shares. As these companies seek more growth, the complexity of crew and route scheduling programs have increased in size and complexity. Scheduling crew, specifically, is one of the major planning problems in the travel industry. To mitigate that, companies have now begun to lease software that use intense mathematical models and algorithms to schedule their crews (Devechi, 2018).

While this has expedited the scheduling process, the argument arises that companies are now treating their crews as data sets instead of human beings. These methods of scheduling are built on optimization, the aim being that the company’s profit outweighs the operation spending. Currently, the algorithms do not account for crew health. This becomes an ethical conundrum. What are the ethical implications of the use of these scheduling optimization algorithms?

Aims

The aim of this project is to:

A) Understand the mathematical models used to schedule crew in the Airline industries

B) Research the physical consequences (positive or negative) that crews undergo due to their current schedules

C) Examine the ethical violations of relying on the methods used for crew scheduling, if any

D) Propose future directions of scheduling algorithms used in this context

Methodology

This project will collect and analyze a mix of quantitative data — in the form of algorithms, routes and schedules used currently, and qualitative data in the form of interviews from crew
members and scheduling analysts from the Airlines. Not only that but we will also be hypothesizing and suggesting ethical recommendation based on collected literature.

**Expected Outcomes**

It is expected that when considering the ethical implications of using optimized algorithms to schedule crew, that violations and conflicts will arise. This is important because, while we cannot stop the growth of the transit industry, we can ensure that later iterations of new software will be designed with the ethical treatment of crew in mind.
Only a hard-deterministic (Adler, 2008) explanation of technology can assume a known outcome of what technologies can offer in terms of social realities and social change. However, the complexities of socio-technical systems imply that the future is unknown; as it is contingent upon the context of technological development, deployment, and systems of use (Winner, 1986). This contingency is at odds with a utopian/dystopian view of autonomous technologies.

Technology is neither utopian, nor dystopian. The social and economic system in which technologies are embedded, determine social evolution and technological impact at any given moment in history. Technologies can shape social visions of enchantment and liberation. But they can also facilitate a social order that is relentlessly harsh, destructive, and miserable for a majority of people (King, 1996). In the late 60’s and early 70’s the idea of AI and autonomous technologies brought hope for social freedom; envisioning societies so advanced that labor and work would be eliminated. The Fun Palace by Cedric Price (Price, 1968) or the Walking Cities by Archigram (Stiener, 2013) are examples of utopian cravings which incentivized rebellious and anarchist movements at the time. On the other hand, communication technologies were critiqued for subjugating societies for spreading consumer culture: mass media was now responsible for class alienation, cultural homogenization, and replacing social relations with mere representation as part of commercialization effects (Debord, 1967). Such contradictory effects of technologies are recurring in the history of social evolution.

As technology becomes more and more intertwined with the society, it is necessary to collaborate with the public to develop and transform technologies for all groups and services in both micro and macro levels. For example, while AI systems provide autonomy and freedom, their development must be questioned to prevent social biases and negative effects in the future. Expanding the possibilities of future autonomous and robotic systems highly relies on understanding and responding to society’s needs, expectations, and mental models toward AI-systems. Imagined affordances of technological artifacts are constantly redefined through the interplay of agency between people and technologies; the quality and context of which can impact the direction of technological development. Therefore, for technologies to reach their full potential- for the better- it is essential to design interactions which are transparent, provide agency to their users, and create trust towards technologies.

Transparency, agency, and trust are the building blocks to create an Ethical AI. However, lack of definitions and loose policies regarding this concept is slowly diminishing reliability and intimacy with AI-systems. Hopefully, a conversation over the dangers of autonomous technological development, and methods to which public debate and protest could affect the design, diffusion, use, and regulation of autonomous technologies, could open paths for developing a more beneficial AI.
Bibliography


Autonomous Tools and Delegation
Jeffrey V. Nickerson
Stevens Institute of Technology

There is little to admire in AI systems whose main goal is surveillance. These systems capture traces of both digital and physical behavior and makes guesses about the future behavior of humans in order to control or take advantage of this behavior [6]. Such AI systems can nudge votes, purchases, and associations between people. They can instigate disruptions in society. They can do this because they can see the networks and collective behaviors that individuals can’t see. That is, aggregated data permit inferences and actions that are beyond the scope of what an individual can see or affect. Larger stores of data and larger compute farms lead to more effective surveillance, which in turn leads to more effective prediction and behavior modification.

Another path that AI is following positions AI as tools rather than as systems. These tools are granted some degree of autonomy, but ultimately are managed by humans, who not only delegate but also monitor the tools. Take, for example, bots in Wikipedia [5]. They generate knowledge. They act on their own. But they have human operators who are responsible for their behavior. In a similar fashion, video game designers use autonomous tools to generate game landscapes and chip designers use autonomous tools to generate layouts [3].

Many of the issues we confront in AI have been studied in management. They involve the functions of the executive. That is, we as humans are building machines that can act autonomously. These machines need to be trained. They can be delegated to, but, as with any delegation situation, there is an obligation to monitor and correct. Principles of management apply [1, 4]. But there are new issues. Humans and machines have different embodiments [2]. We need new modes of delegation and monitoring to take into account differences in the speed and nature of processing that takes place in autonomous tools and their managers. For example, very fast decision-makers need very fast monitors, and these monitors may need to be machines. Thus managers become collectives of humans and machines.

References

Innovation in the artificial intelligence space has been propelled by an ethos of progress and growth at the expense of the consideration of ethics, democratic values and human rights. As companies like Facebook, Twitter and YouTube have expanded their use of machine learning (ML) algorithms in prioritizing information for users, but also in allowing programmers to interact with APIs by way of personally generated AI-enabled software, they have opened up global publics to automated manipulation and surveillance. Both technology firms and governments have failed to effectively protect user’s privacy in an era where individual ad targeting—scaffolded by AI software—has extended to spheres political and personal. Meanwhile, AI and ML are presented by tech executives, including Mark Zuckerberg in his hearings before congress, as a cure-all “McGuffin” to the problems associated with mis- and dis-information and state-sponsored computational propaganda. Undoubtedly, AI will play a crucial future role in detecting and mitigating these problems at scale. At present, however it is unclear how or whether these systems will serve to protect rather than transgress.

Alongside game designer and author Jane McGonigal, and with the support and collaboration of Omidyar Networks and the Institute for the Future, I have designed the Ethical Operating System (ethicalos.org). The Ethical OS is a guide to anticipating the future impact of today’s technology. In building this toolkit, we specifically focused upon technology designers and questions at the heart of innovation in AI. We identify eight risk zones (see graphic below) associated with the current state of design thinking in tech sector: 1) truth, disinformation, propaganda, 2) addiction and the dopamine economy, 3) economic and asset inequalities, 4) machine ethics and algorithmic biases, 5) surveillance state, 6) data control and monetization, 7) implicit trust and user understandings, and 8) hateful and criminal actors. Undoubtedly, there is cross-over between these groups but we feel that the risk zones together, along with the larger Ethical OS framework, provide a starting point for asking crucial questions when building new AI technology or amending old systems.

I believe the insights developed while constructing the Ethical OS, as well as ideas from my broader work on computational propaganda and emergent technologies, would be useful and provocative additions to the Good Systems: Ethical AI for CSCW workshop. The Ethical OS Toolkit is now being used by numerous large tech firms, start-ups, incubators and venture-capital firms. It has been taught in computer and information science courses at Stanford, UC Berkeley and the University of Texas at Austin as well as through MOOC courses (also through Stanford). I believe what we’ve learned in the course of our active and ongoing research will serve to undergird the insights of other researchers at this CSCW workshop. Ultimately, I hope to use this toolkit and my prior research to argue that the future of AI can be made more hopeful if we begin to design technology with human rights—and other tenants of democracy—in mind today.
Risk Zones to Consider When Designing and Launching New Socio-Technological Systems

Image Credit: Omidyar Networks and the Institute for the Future w/ Jane McGonigal and Samuel Woolley.