AI Matters

Annotated Table of Contents

Links

SIGAI website: http://sigai.acm.org/
Newsletter: http://sigai.acm.org/aimatters/
Blog: http://sigai.acm.org/ai-matters/
Twitter: http://twitter.com/acm_sigai/
Edition DOI: 10.1145/3340470

Join SIGAI

Students $11, others $25
For details, see http://sigai.acm.org/
Benefits: regular, student

Also consider joining ACM.

Our mailing list is open to all.

Notice to Contributing Authors to SIG Newsletters

By submitting your article for distribution in this Special Interest Group publication, you hereby grant to ACM the following non-exclusive, perpetual, worldwide rights:

- to publish in print on condition of acceptance by the editor
- to digitize and post your article in the electronic version of this publication
- to include the article in the ACM Digital Library and in any Digital Library related services
- to allow users to make a personal copy of the article for noncommercial, educational or research purposes

However, as a contributing author, you retain copyright to your article and ACM will refer requests for republication directly to you.

Submit to AI Matters!

We're accepting articles and announcements now for the next issue. Details on the submission process are available at http://sigai.acm.org/aimatters.

AI Matters Editorial Board

Contact us: aimatters@sigai.acm.org

Contents Legend

📗 Book Announcement
📕 Ph.D. Dissertation Briefing
🎓 AI Education
📅 Event Report
🔥 Hot Topics
🎭 Humor
❗ AI Impact
📰 AI News
💬 Opinion
📄 Paper Précis
🔦 Spotlight
📷 Video or Image

Details at http://sigai.acm.org/aimatters

Welcome to AI Matters 5(2)

Amy McGovern, co-editor (University of Oklahoma; aimatters@sigai.acm.org)
Iolanda Leite, co-editor (Royal Institute of Technology (KTH); aimatters@sigai.acm.org)

DOI: 10.1145/3340470.3340471

Issue overview

Welcome to the second issue of the fifth volume of the AI Matters Newsletter. We have exciting news from SIGAI Vice-Chair Sanmay Das: ”We are delighted to announce that the first ever ACM SIGAI Industry Award for Excellence in Artificial Intelligence will be awarded to the Decision Service created by the Real World Reinforcement Learning Team from Microsoft! The award will be presented at IJCAI 2019. For more on the award and the team that received it, please see https://sigai.acm.org/awards/industry_award.html.”

This issue opens with our interview series, where Marion Neumann interviews Leslie Pack Kaelbling, a Professor of Computer Science and Engineering at MIT and founder of the Journal of Machine Learning Research.

In our regular columns, we have a summary of upcoming AI conferences and events from Michael Rovatsos. Todd Neller's educational column is dedicated to a *Magic: The Gathering* dataset that provides interesting opportunities for exploring research questions on data science and ML. In the policy column, Larry Medsker summarizes recent AI related initiatives and discusses new jobs in the AI future.

This issue features the first set of winning essays from the 2018 ACM SIGAI Student Essay Contest, with the second set of winning essays to appear in the next issue. In addition to having their essay appear in AI Matters, the contest winners received either monetary prizes or one-on-one Skype sessions with leading AI researchers.

In the regular contributed papers, our editors Cameron Hughes and Tracey Hughes propose potential metrics for commercial AI.

We close with our (now regular) entertainment column, an AI generated crossword puzzle by Adi Botea. You can also find the solution to the puzzle from the previous issue.

Submit to AI Matters!

Thanks for reading! Don't forget to send your ideas and future submissions to *AI Matters*! We're accepting articles and announcements now for the next issue. Details on the submission process are available at http://sigai.acm.org/aimatters.

Amy McGovern is co-editor of AI Matters. She is a Professor of computer science at the University of Oklahoma and an adjunct Professor of meteorology. She directs the Interaction, Discovery, Exploration and Adaptation (IDEA) lab. Her research focuses on machine learning and data mining with applications to high-impact weather.

Iolanda Leite is co-editor of AI Matters. She is an Assistant Professor at the School of Electrical Engineering and Computer Science at the KTH Royal Institute of Technology in Sweden. Her research interests are in the areas of Human-Robot Interaction and Artificial Intelligence. She aims to develop autonomous socially intelligent robots that can assist people over long periods of time.

AI Profiles: An Interview with Leslie Kaelbling

Marion Neumann (Washington University in St. Louis; m.neumann@wustl.edu)

DOI: 10.1145/3340470.3340472

Introduction

Welcome to the eighth interview profiling a senior AI researcher. This time we will hear from Leslie Kaelbling, Panasonic Professor of Computer Science and Engineering in the Department of Electrical Engineering and Computer Science at MIT.

Figure 1: Leslie Kaelbling

Biography

Leslie is a Professor at MIT. She has an undergraduate degree in Philosophy and a PhD in Computer Science from Stanford, and was previously on the faculty at Brown University. She was the founding editor-in-chief of the Journal of Machine Learning Research. Her research agenda is to make intelligent robots using methods including estimation, learning, planning, and reasoning. She is not a robot.

Getting to Know Leslie Kaelbling

When and how did you become interested in CS and AI?

I went to high school in rural California, but the summer before my senior year I went to an NSF summer program in math. We actually ended up studying computer science. My crowning achievement was writing quicksort in Basic! I also discovered Scientific American, and started reading Martin Gardner's columns (the only part of the magazine I could even sort of understand) and learned about *Gödel, Escher, Bach* by Douglas Hofstadter. I managed to get a copy of it, and that's what made me really get interested in AI.

What professional achievement are you most proud of?

Starting JMLR, I guess. I think it's been very helpful for the community, and was actually not very hard to do.

What would you have chosen as your career if you hadn't gone into CS?

No idea! I'm pretty flexible. But almost sure something technical.

What is the most interesting project you are currently involved with?

I'm doing the same thing I've always been doing, which is trying to figure out how to make really intelligent robots. I do this mostly out of curiosity: I want to understand what the necessary and sufficient computational methods are for making an agent that behaves in a way we'd all be happy to call intelligent. I think human intelligence is probably a point in a big space of computational mechanisms that achieve intelligent behavior. I'm interested in understanding that whole space!

AI is grown up – it's time to make use of it for good. Which real-world problem would you like to see solved by AI in the future?

I'm not so focused on solving actual problems, but I'm fairly sure that methods that are developed on the way to understanding computational approaches to intelligent behavior will end up being useful in a variety of ways that I don't anticipate.

How can we make AI more diverse? Do you have a concrete idea on what we as (PhD) students, researchers, and educators in AI can do to increase diversity our field?

Unfortunately, I don't, really. The answer for AI is probably not substantially different from the answer for CS or even engineering more broadly.

How do you balance being involved in so many different aspects of the AI community?

I'm a good juggler! But it's suddenly much harder than it was, just because of the enormous growth of enthusiasm about AI, and machine learning in particular. Everything I do, from teaching undergraduates to graduate admissions to hiring to writing tenure letters to reviewing papers to organizing conferences has just gotten an order of magnitude bigger and more complex. I was really affected by this for a while, but now I'm honing my "no"-saying skills so I can protect time to actually do research (which is why I'm in this business, after all).

What do you wish you had known as a Ph.D. student or early researcher?

I don't know. Things worked out pretty well for me, but completely by accident. I think there are many ways in which it's actually good to not know much. You have a greater chance of doing something really novel or really hard just because you don't know it's novel or hard.

What is your favorite AI-related movie or book and why?

Well, *Gödel, Escher, Bach* was formative for me. Its focus on primitives and systems of combination, and themes of recursion, semantics, quotation, reflection really resonated with me and I'm sure that the ways in which I think and formulate problems still show its influences. I haven't re-read it since I was 17, though, so I don't know what it would feel like now.

Help us determine who should be in the AI Matters spotlight!

If you have suggestions for who we should profile next, please feel free to contact us via email at aimatters@sigai.acm.org.

Events

Michael Rovatsos (University of Edinburgh; mrovatso@inf.ed.ac.uk)

DOI: 10.1145/3340470.3340473

This section features information about upcoming events relevant to the readers of AI Matters, including those supported by SIGAI. We would love to hear from you if you are are organizing an event and would be interested in cooperating with SIGAI, or if you have announcements relevant to SIGAI. For more information about conference support visit sigai.acm.org/activities/requesting_sponsorship.html.

17th International Conference on Artificial Intelligence and Law (ICAIL'19)

Montreal, Canada, June 17-20, 2019
https://icail2019-cyberjustice.com
The 2019 edition of the International Conference on Artificial Intelligence and Law (ICAIL) will be held at the Cyberjustice Laboratory, University of Montreal.The conference is held biennially under the auspices of the International Association for Artificial Intelligence and Law (IAAIL) and in cooperation with the Association for the Advancement of Artificial Intelligence (AAAI) and the ACM Special Interest Group on Artificial Intelligence (ACM SIGAI). The conference proceedings are published by ACM.

14th International Conference on the Foundations of Digital Games (FDG'19)

San Louis Obispo, CA, August 26-30, 2019
fdg2019.org
Foundations of Digital Games is a major international "big tent" academic conference dedicated to exploring the latest research in all aspects of digital games. FDG is usually held in Europe or North America once a year. FDG 2018 was held in Malmo, Sweden. FDG 2019 is held in cooperation with ACM and ACM SIG AI, SIGGRAPH and SIGCHI.

11th International Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K'19)

Vienna, Austria, September 17-19, 2019
www.ic3k.org
The purpose of IC3K is to bring together researchers, engineers and practitioners on the areas of Knowledge Discovery, Knowledge Engineering and Knowledge Management. IC3K is composed of three co-located conferences, each specialized in at least one of the aforementioned main knowledge areas. IC3K 2019 will be held in conjunction with WEBIST 2019 and IJCCI 2019.
Submission deadline: June 12, 2019

11th International Joint Conference on Computational Intelligence (IJCCI'19)

Vienna, Austria, September 17-19, 2019
http://www.ijcci.org
The purpose of IJCCI is to bring together researchers, engineers and practitioners on the areas of Fuzzy Computation, Evolutionary Computation and Neural Computation. IJCCI is composed of three co-located conferences, each specialized in at least one of the aforementioned main knowledge areas. IJCCI 2019 will be held in conjunction with WEBIST 2019 and IC3K 2019.
Submission deadline: June 12, 2019

International Conference on Web Intelligence (WI'19)

Thessaloniki, Greece, October 14-17, 2019
https://webintelligence2019.com
Web Intelligence (WI) aims to achieve a multi-disciplinary balance between research advances in the fields of collective intelligence, data science, human-centric computing, knowledge management, and network science. It is committed to addressing research that deepens the understanding of computational, logical, cognitive, physical as well as business and social foundations of the future Web, and enables the development and

application of intelligent technologies. WI ’19 features high-quality, original research papers and real-world applications in all theoretical and technological areas that make up the field of WI.

25th International Conference on Intelligent User Interfaces (IUI’20)

Cagliari, Italy, March 17-20, 2020
https://iui.acm.org/2020/
ACM IUI 2020 is the 25th annual meeting of the intelligent interfaces community and serves as a premier international forum for reporting outstanding research and development on intelligent user interfaces. ACM IUI is where the Human-Computer Interaction (HCI) community meets the Artificial Intelligence (AI) community. We are also very interested in contributions from related fields, such as psychology, behavioral science, cognitive science, computer graphics, design, the arts.
Submission deadline: 8th October 2019

Michael Rovatsos is the Conference Coordination Officer for ACM SIGAI, and a faculty member at the University of Edinburgh. His research is in multiagent systems and human-friendly AI. Contact him at mrovatso@inf.ed.ac.uk.

AI Education Matters: Data Science and Machine Learning with Magic: The Gathering

Todd W. Neller (Gettysburg College; tneller@gettysburg.edu)

DOI: 10.1145/3340470.3340474

Introduction

In this column, we briefly describe a rich dataset with many opportunities for interesting data science and machine learning assignments and research projects, we take up a simple question, and we offer code illustrating use of the dataset in pursuit of answers to the question.

Magic: The Gathering (MTG) is a collectible card game featuring imperfect information, chance, complex hand management, and fascinating metagame economics in the online market for cards. One can study the in-game economy that concerns different colored/uncolored *mana*, or one can study the time-series dollar costs of cards on the open market as they are introduced and then fluctuate in supply and demand, e.g. when rotating out of standard play legality (generally after less than two years), or when found to have powerful uses in newer play formats like Pauper.

As of 25 February 2019, a freely-available, rich JSON card dataset for 19386 MTG cards may be downloaded from the MTGJSON website. For this column, we select a single focus question: How much more is the in-game cost for a creature with “flying” ability.

The Cost of Flying

Flying is an evasive ability of creatures that generally comes with a higher converted mana cost (CMC), i.e. the total number of in-game currency units needed to bring the card into play regardless of the specific mana colors required.

In related work, Henning Hasemann, a Berlin-based researcher and software engineer, has done preliminary data science and machine learning investigation on the cost of flying and other keyword attributes of MTG cards[1]. He concluded that, for the same power and toughness, one pays approximately 0.79 in CMC for flying. We will take a slightly different approach in this column.

Having downloaded AllCards.json, one can easily load the card data in Python using Python’s native `json` library:

```
import json
with open(’AllCards.json’, ’r’,
    encoding=’utf8’) as read_file:
    data = json.load(read_file)
print(len(data), ’cards read.’)
```

One can then iterate through cards and their respective data dictionaries:

```
for card_name in data:
    card_data = data[card_name]
```

For each card dictionary `card_data`, we can access rule text, power, toughness, and CMC through dictionary keys `’text’`, `’power’`, `’toughness’`, and `’convertedManaCost’`. The structure of the JSON data is well-documented.

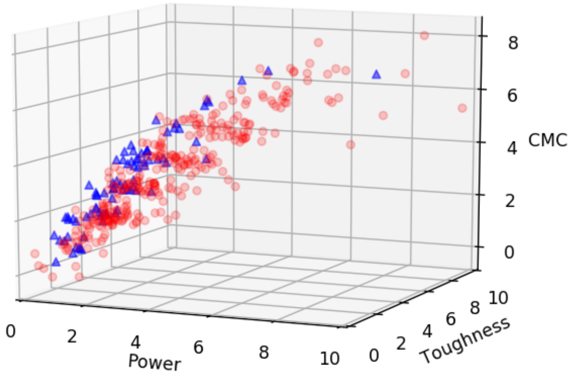
More difficult to ascertain is which cards are creatures which are inherently “flying”. Some non-creature cards contain rule text affecting flying creatures. Other creature cards describe defense against flying creatures. A simple approach to eliminate false positive and false negatives is to filter for only creatures that have either no rule text or rule text consisting of “Flying” only. Looking over the data filtered thus, we noted that cards with “transform” layout should be excluded as their true cost of flying is the mana cost of the initial side plus the game condition that must be achieved to flip the card to its transformed side.

Instructors and students can jumpstart their exploration of the following overview by downloading all data and code, including the aforementioned AllCards.json data, Python code, and a corresponding Jupyter notebook.

In both the code and the Jupyter notebook, one can trace a simple exploration of this

[1] http://leetless.de/tag-MTG.html

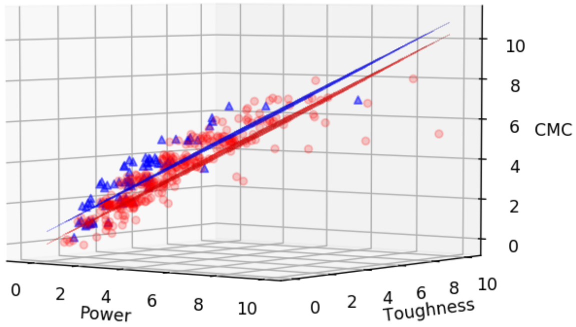
question where, having filtered the 9844 creatures conservatively to include only 81 flying and 312 non-flying creatures, we first view the data as a jittered scatterplot where flying and non-flying creatures are represented as blue triangles and red circles, respectively:



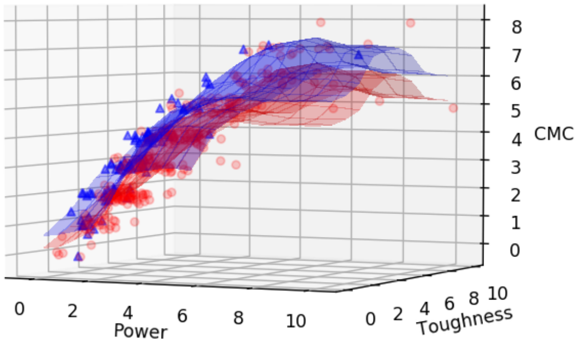
Starting simply, we perform and visualize a linear regression predicting CMC from power, toughness, and a binary "flying" attribute, yielding a linear model with an R-squared value of 0.82:

$$\begin{aligned} \text{CMC} &= 0.6111629366158998 * \text{power} \\ &+ 0.41265041398361274 * \text{toughness} \\ &+ 0.60820253825056 * \text{flying} \\ &+ 0.2782416745722158 \end{aligned}$$

Thus, this linear model would predict a flying premium of about 0.61 CMC. Here we visualize the predictions for flying and non-flying creatures as blue and red planes, respectively:



Looking along the planes, we see non-linearity in the data. Applying gradient boosted decision tree learning (sklearn's GradientBoostingRegressor), we see a more nuanced prediction:



A sample-weighted mean of the flying premium with this model is about 0.725 CMC, and one can observe that the flying premium increases to about 1 for toughness values 4 and greater which lead to more difficult creature removal for the opponent.

However, the greater takeaway here is that this free dataset is rich and complex as the game, inviting simple inquiries like this and offering opportunities for much more complex analyses. The example code offered here is an invitation for educators and their students to take up a wide variety of fun and interesting questions concerning this popular 25-year-old game.

Past and Future MTG Research

At present, published AI research on MTG is limited to little more than a few papers on Monte Carlo Tree Search application to a simplified MTG card set (Ward & Cowling, 2009; Cowling, Ward, & Powley, 2012) and procedural MTG card generation (Summerville & Mateas, 2016). Hearthstone, a very similar digital collectible card game, has received much more research attention.

MTG should be of great future interest as a general game play challenge for a variety of reasons beyond its popularity and 25-year history. First and foremost is the greater strategic complexity arising from the rules of over nineteen thousand MTG cards. The software engineering challenge of modeling MTG is likely the reason more attention has been given to the newer and simpler game Hearthstone. An important building-block challenge here would be to design AI systems that could learn to play two fixed decks optimally against one another.

A second reason for interest is the fascinating metagame choices where players build decks for competition at time of play (e.g. through drafting) or a priori (e.g. through constructed formats) that seek to find beneficial synergy between a constrained set of allowable cards. A clever recent construction showed MTG to be Turing complete (Churchill, Biderman, & Herrick, 2019).

A third reason could be termed the *metametagame*, where the MTG consumer negotiates the supply and demand cost dynamics controlled by Wizards of the Coast LLC (WotC) by creating new play formats (e.g. Pauper, Cube) to constrain costs. In his 2016 GDC talk, MTG head designer Mark Rosewater's lesson 18 was that "restrictions breed creativity". Over the history of MTG, players have shown that *cost* restrictions breed *affordable* creativity. For example, Cube designers seek to select sets of 360 or more unique MTG cards that allow interesting gameplay for players that repeatedly draft from the Cube.

Whereas deck construction could be said to invite players to become designers of the games they play, play format design could be said to invite players to become metagame designers. When AI systems of the future are capable of such metametalevel decisions in this imperfect information game of chance, we will see an unprecedented leap forward in game design itself.

In our modest present time, we hope that you and your students find interesting simple research questions to explore in the MTGJSON dataset and sow seeds of curiosity for research advances to come.

References

Churchill, A., Biderman, S., & Herrick, A. (2019). Magic: The gathering is turing complete. *CoRR*, *abs/1904.09828*. Retrieved from http://arxiv.org/abs/1904.09828

Cowling, P., Ward, C., & Powley, E. (2012). Ensemble determinization in Monte Carlo tree search for the imperfect information card game "Magic: The Gathering". *IEEE Transactions on Computational Intelligence and AI in Games*, *4*(4), 241–257. doi: 10.1109/TCIAIG.2012.2204883

Summerville, A., & Mateas, M. (2016). Mystical tutor: A Magic: The Gathering design assistant via denoising sequence-to-sequence learning. In *Proc. AAAI conference on artificial intelligence and interactive digital entertainment 2016.*

Ward, C., & Cowling, P. (2009). Monte Carlo search applied to card selection in Magic: The Gathering. In *Proc. 2009 IEEE symposium on computational intelligence and games* (pp. 9–16). IEEE. doi: 10.1109/CIG.2009.5286501

Todd W. Neller is a Professor of Computer Science at Gettysburg College. A game enthusiast, Neller researches game AI techniques and their uses in undergraduate education.

AI Policy Matters

Larry Medsker (The George Washington University; lrm@gwu.edu)

DOI: 10.1145/3340470.3340475

Abstract

AI Policy is a regular column in *AI Matters* featuring summaries and commentary based on postings that appear twice a month in the *AI Matters* blog (`https://sigai.acm.org/aimatters/blog/`). We welcome everyone to make blog comments so we can develop a rich knowledge base of information and ideas representing the SIGAI members.

News and Announcements

AAAI Policy Initiative

AAAI has established a new mailing list on US Policy that will focus exclusively on the discussion of US policy matters related to artificial intelligence. All members and affiliates are invited to join the list at this link. Participants will have the opportunity to subscribe or unsubscribe at any time. The mailing list will be moderated, and all posts will be approved before dissemination. This is a great opportunity for another productive partnership between AAAI and SIGAI for policy work.

EPIC Panel on June 5th

A panel on AI, Human Rights, and US policy, was hosted by the Electronic Privacy Information Center (EPIC) at their annual meeting (and celebration of 25th anniversary) on June 5, 2019, at the National Press Club in DC. Lorraine Kisselburgh (Purdue) joined Harry Lewis (Harvard), Sherry Turkle (MIT), Lynne Parker (UTenn and White House OSTP director for AI), Sarah Box (OECD), and Bilyana Petkova (EPIC and Maastricht) to discuss AI policy directions for the US.

AI Research Roadmap

The Computing Community Consortium (CCC) received comments on the draft of *A 20-Year Community Roadmap for AI Research in the US*. The draft was the result of a community process involving more than one hundred AI professionals. The CCC initiative to create the Roadmap for Artificial Intelligence was started in Fall, 2018, under the leadership of Yolanda Gil (University of Southern California and President of AAAI) and Bart Selman (Cornell University and President Elect of AAAI). Follow this link to the whole report.

New Jobs in the AI Future

As employers increasingly adopt automation technology, many workforce analysts look to jobs and career paths in new disciplines, especially data science and applications of AI, to absorb workers who are displaced by automation. By some accounts, data science is in first place for technology career opportunities. Estimating current and near-term numbers of data scientists and AI professionals is difficult because of different job titles and position descriptions used by organizations and job recruiters. Likewise, many employees in positions with traditional titles have transitioned to data science and AI work. Better estimates, and at least upper limits, are necessary for evidence-based predictions of unemployment rates due to automation over the next decade.

McKinsey & Company estimates 375 million jobs will be lost globally due to AI and other automation technologies by 2030, and one school of thought in today's public discourse is that at least that number of new jobs will be created. An issue for the AI community and policy makers is the nature, quality, and number of the new jobs – and how many data science and AI technology jobs will contribute to meeting the shortfall.

An article in *KDnuggets* by Gregory Piatetsky points out that a “Search for data scientist (without quotes) finds about 30,000 jobs, but we are not sure how many of those jobs are for scientists in other areas – persons employed to analyze and interpret complex digital data, such as the usage statistics of a website, especially in order to assist a business in its decision-making. Titles include Data Scientist, Data Analyst , Statistician, Bioinformatician, Neuroscientist, Marketing executive, Computer scientist, etc.”

Data on this issue could clarify the net number of future jobs in AI, data science, and related areas. Computer science had a similar history with the boom in the new field followed by migration of computing into many other disciplines. Another factor is that “long-term, however, automation will be replacing many jobs in the industry, and Data Scientist jobs will not be an exception. Already today companies like DataRobot and H2O offer automated solutions to Data Science problems. Respondents to a *KDnuggets* 2015 Poll expected that most expert-level Predictive Analytics and Data Science tasks will be automated by 2025. To stay employed, Data Scientists should focus on developing skills that are harder to automate, like business understanding, explanation, and story telling.” This issue is important in estimating the number of new jobs by 2030 for displaced workers.

Kiran Garimella in his Forbes article “Job Loss From AI? There’s More To Fear!” examines the scenario of not enough new jobs to replace ones lost through automation. His interesting perspective turns to economists, sociologists, and insightful policymakers “to reexamine and re-formulate their models of human interaction and organization and ... rethink incentives and agency relationships”.

How Open Source?

A recent controversy erupted over OpenAI’s new version of their language model for generating well-written next words of text based on unsupervised analysis of large samples of writing. Their announcement and decision not to follow open-source practices raises interesting policy issues about regulation and self-regulation of AI products. OpenAI, a nonprofit AI research company founded by Elon Musk and others, announced on February 14, 2019, that “We’ve trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization – all without task-specific training”.

The reactions to the announcement followed from the decision behind the following statement in the release: “Due to our concerns about malicious applications of the technology, we are not releasing the trained model. As an experiment in responsible disclosure, we are instead releasing a much smaller model for researchers to experiment with, as well as a technical paper”.

The Electronic Frontier Foundation has an analysis of the manner of the release (letting journalists know first) and concludes, “when an otherwise respected research entity like OpenAI makes a unilateral decision to go against the trend of full release, it endangers the open publication norms that currently prevail in language understanding research”.

This issue is an example of previous ideas in our Public Policy blog about who, if anyone, should regulate AI developments and products that have potential negative impacts on society. Do we rely on self-regulation or require governmental regulations? What if the U.S. has regulations and other countries do not? Would a clearinghouse approach put profit-based pressure on developers and corporations? Can the open source movement be successful without regulatory assistance?

Please join our discussions at the SIGAI Policy Blog.

Larry Medsker is Research Professor of Physics and was founding director of the Data Science graduate program at The George Washington University. He is a faculty member in the GW Human-Technology Collaboration Lab and Ph.D. program. His research in AI includes work on artificial neural networks, hybrid intelligent systems, and the impacts of AI on society and policy. He is the Public Policy Officer for the ACM SIGAI.

Beyond Transparency: A Proposed Framework for Accountability in Decision-Making AI Systems

Janelle Berscheid (University of Saskatchewan; j.berscheid@usask.ca)
Francois Roewer-Despres (University of Toronto; francoisrd@cs.toronto.edu)

DOI: 10.1145/3340470.3340476

Abstract

Transparency in decision-making AI systems can only become actionable in practice when all stakeholders share responsibility for validating outcomes. We propose a three-party regulatory framework that incentivizes collaborative development in the AI ecosystem and guarantees fairness and accountability are not merely afterthoughts in high-impact domains.

Introduction

Decision-making AI systems are becoming commonplace due to recent and rapid advances in computer hardware, machine learning algorithms, and an explosion in data availability. Increasingly, AI is being deployed to make life-altering decisions, such as those involving health, freedom, security, finance, and livelihood. However, the public is apprehensive about automated decision-making: the majority of Americans in a Pew Research Center survey were opposed to current applications of decision-making systems (Smith, 2018). Chief among their worries were concerns regarding the potential bias and unfairness of these systems, as well as skepticism regarding the validity of the decisions being made. Research on the social impact of AI, such as from New York University's AI Now Institute, corroborates these concerns, arguing that automated decision-making systems are often deployed untested and without accountability measures or processes for appeal (Whittaker et al., 2018).

Such technologies often create a gap between private and social cost. The motivating example for this essay was the COMPAS recidivism algorithm, developed by a company named Northpointe, which came under fire after a ProPublica investigation found that its results were racially biased and that jurisdictions were misapplying the results during trials, resulting in defendants incorrectly being labeled as potentially violent re-offenders (Angwin & Larson, 2016).

Stifling innovation is undesirable, but unchecked deployment of high-impact decision-making systems is damaging to the long-term health of the AI ecosystem. The public backlash will likely only intensify as the inevitable failures of unvetted systems come to light. Properly vetted systems have the potential to save time, money, and even lives, so finding remedies to the negative aspects of AI is critical to ensuring the public is a stakeholder in the AI ecosystem rather than an adversary.

Properly-designed AI systems present a unique opportunity to make transparent decisions by circumventing human fallibility. In fact, the framework we are proposing holds AI decision-makers to far more stringent standards than could ever be applied to human decision-makers. For example, psychological studies demonstrate human tendencies to make decisions subconsciously and then consciously rationalizing them post-hoc (Soon, Brass, Heinze, & Haynes, 2008), or to be prefer one decision over another based on framing of the situation (Tversky & Kahneman, 1981).

Ensuring accountability when mistakes are made is step towards addressing the concerns surrounding bias, fairness, and validity. Transparent decision-making—where the reasoning behind an AI's decision is made clear, interpretable, and auditable—is often proposed as a solution to the problem of biased or invalid AI systems. Indeed, transparency addresses many of these concerns; however, the gap between AI developers and other parties, such as the public and policymakers, cannot be closed by transparency alone. For example, the recent testimony of Facebook CEO Mark Zuckerberg before

Congress regarding the exploitation of Facebook users' data increased awareness surrounding this data misuse, but also revealed that many Congresspeople lack even a basic understanding of the technology (Kang, Kaplan, & Fandos, 2018). As a result, many pertinent questions were not asked, leading many to view it as a missed opportunity (Freedland, 2018).

Thus, we argue that transparency alone is not a sufficient requirement to produce accountability and reassure stakeholders: additional elements are needed to make transparency actionable in practice. For the public to become comfortable with autonomous decision-making systems, there needs to be a sense that a person can take remedial actions against an AI system if wronged. We identify two requirements to meeting such demands: disclosure of the existence of an AI in a decision-making process to the public, coupled with an appeal process for the AI's decisions, and a validated scope for the AI, including the use cases for which it been tested and its limitations.

These components suggest a need for a closer relationship between those who develop AI systems and those who create policies for their deployment. We propose a regulatory framework for AI systems that captures this need for communication into the development and deployment process itself. We envision establishing a pair of documents that ensure these disclosure and scope requirements are fulfilled prior. The developer's document publicly declares the scope of their system and how it has been validated, while the client document explicates the disclosure and appeal policies surrounding the deployment of a developer's system into a particular domain. In addition, a formally-established relationship between the developers and clients helps reduce cases where these two parties try to pin a system's failure on the other. In turn, this would facilitate the open sharing of knowledge and the cooperative development of fair procedures.

Definitions

For the purposes of this article, we want to make a careful distinction between the parties we have termed "developer", "client", and "subject".

Developer
An entity or person that has created an AI system for the purpose of real-world deployment. This may include, but is not limited to, those who designed the system, those who trained and tested the model at the core of the system, and those who developed the data pipeline.

Client
An entity or person that intends to deploy a developer's AI system as part of a decision-making process. This process may include other human or automated elements, as well as integrate AI systems from multiple developers. Typically, the client will be an institution (such as a charity, hospital, or government body) or a private company.

Subject
An entity or person that is the target of a client's deployed AI system. Subjects include patients, inmates, and consumers, or even private companies.

This distinction reflects a separation of concerns: developers are primarily concerned with the technical components of a system; clients are primarily concerned with the use of a system in relation to the subjects for a specific context. Further, this separation allows a developer to license their system to more than one client. It remains possible for the developer and client to be the same entity.

In addition, we want to clarify the types of AI systems being considered.

AI System
A system or process that incorporates AI in some form or another for the purpose of decision-making.

In this article, we will be considering requirements *only* for decision-making AI systems. We not do consider any AI systems that interact with, or perhaps mimic humans, as these may have a different set of requirements. We also focus our discussion on systems in high-impact areas, which may include, but are not limited to, health services, financial institutions, justice systems, aid programs, and civic initiatives. AI systems in lower-impact areas, such as recommender systems for consumer platforms, are less urgently in need of regulation, though these systems could benefit from the framework we describe as well. We also exempt AI research from our discussion, instead choosing to focus on systems intended for real-world deployment.

Transparency

We define transparency as making a decision that is in some way explainable, that has its inner workings available for review, and is not proprietary. This differs from disclosure, which is 'transparency in the use of a model' (rather than 'transparency in the prediction of a model'). In other words, the use of a model can be disclosed, but the model itself might not be interpretable.

Implemented properly, transparency can help us examine erroneous assumptions made about the data, and even let human operators correct bad features in the model (Ribeiro, Singh, & Guestrin, 2016). Transparency also allows us to examine current and possible points of failure in a system, as well as monitor system performance to ensure the quality of decision-making does not degrade over time. In addition, transparency can help us verify that public stakeholders' criteria, such as fairness or the removal of demographic bias, are being met.

Currently, many decision-making systems are neither transparent nor easily interpretable. Deep learning techniques, in particular, spur this concern, since neural networks are often characterized as a black box whose decision-making process is completely opaque. Explainable AI (XAI) research seeks to develop techniques to shine light on these systems, yet many fundamental problems remain open. Even the definitions of terms such as "transparency", "interpretability", and "explainability" are difficult to establish and even harder to unify across different input domains.

In addition, the mechanisms by which humans interpret these explanations, especially in relation to natural human subjectivity and bias, has yet to be understood. A recent survey of submissions from the International Joint Conferences on Artificial Intelligence workshop in XAI and found that knowledge from fields such as behavioural research and social science were rarely referenced, and the evaluations of any explanations that were produced did not generally include behavioural experiments (Miller, Howe, & Sonenberg, 2017). Behavioural analyses may be necessary to address subtler issues related to transparency. For example, the adequacy of a particular type of explanation may be dependent on the human interpreter. Indeed, doctors and patients use different explanatory models when interpreting a medical decision (Good, 1993).

Furthermore, hidden feedback loops, where multiple automated decision-making systems influence each other's inputs over time, may also interfere with the system's decision-making in a way that is difficult for a transparency algorithm to identify (Sculley et al., 2015). Transparency in this context may reveal a change over time, but cannot identify the influence of the other systems on the data collected, and thus on the model's decision-making ability.

Despite these unsolved issues, transparency remains an integral component for accountability in AI. Exposing the inner workings of a model to external review helps foster trust in decision-making systems. However, transparency is not directly actionable in deployed systems unless an additional framework is in place to ensure subjects can use transparent explanations to hold developers and clients accountable when wronged.

Disclosure

For AI systems to be transparent, their use must be made known, rather than remaining a hidden part of a decision-making process. Applications that fail to disclose the use of AI systems create a power imbalance, where those unaware of its use are not in a position to question or challenge the figures making crucial decisions on their behalf. This leads to hidden biases and a lack of accountability, as well as creating confusion when problems do arise with these automated systems (Diallo, 2018).

Therefore, we argue that a person significantly impacted by a decision-making system should have the presence of this system clearly disclosed to them. This disclosure could be verbal or—in formal cases—form-based, requiring acknowledgement on the part of the subject. Many application domains of decision-making systems already require paperwork, such as health or legal systems, so viable disclosure channels already exist. A distinction can be made in disclosure between AI-*made* decisions, where the system makes a ruling without any human oversight, and AI-*assisted* decisions, where the system makes recommendations that are reviewed by a human as part of the final decision. Making this distinction could help diagnose points of failure in a decision-making process. First, whenever an AI-assisted process leads to biased or unfair decisions, measures such as retraining the human reviewers may be part of the solution. Second, an AI-assisted decision-making process in which the human reviewer always defers to the AI system's recommendation, may be treated as an effectively AI-made decision, which may violate the disclosure agreement.

In addition, disclosure will help developers identify stakeholders in the decision-making process who may otherwise be overlooked. For example, developers of a recidivism algorithm may think to consult prisoners in their requirements gathering, but may neglect to consult at-risk communities, which ought be consulted to create a fair and unbiased development process. Mandatory disclosure of AI use in their software would create an opportunity for these forgotten stakeholders to make their voices heard through an appeal process, and would encourage developers to expand their definition of impacted communities and stakeholders in future developments, creating a more community-conscious AI development process.

As alluded to, disclosure goes hand-in-hand with the additional requirement that autonomous decisions are appealable. An appeal process that is obscured or hidden cannot be described as fair, as it creates barriers limiting the participation of wronged subjects. When subjected to autonomous decisions, people ought to know not only that AI is employed, but also how its decisions can be examined or appealed. Specifying the details of the appeal process is outside the purview of this article, but we suggest that the process should be inclusive to all affected subjects, and have a minimal burden of entry so that members of marginalized communities could reasonably be expected to go through with the process.

Undoubtedly, this disclosure process would cause friction in the adoption of AI applications by the public. Affected individuals would likely challenge decisions, ask questions, and make appeals more frequently when the presence of AI in a decision-making process is publicized. This should be viewed as a positive long-term effect, rather than as a barrier to innovation and progress: overlooked stakeholders in these systems will be able to engage in a dialogue with the developers, clients, and institutions. Because the long-term health of AI-enabled technologies will ultimately depend on the public's trust and acceptance of these systems, AI developers should be responsible for winning the confidence of both regulating bodies and the general public regarding the efficacy, safety, and fairness of their systems. This will enable a more iterative and democratic process in AI development and deployment.

Validated Scope

Clearly defining the scope of autonomous decision-making systems protects subjects against spurious claims made by developers and clients, and protects developers against misapplication or misuse of their systems by clients. In addition, clearly-defined scope would help narrow the discovery phase, as well as reduce the decision-making burden, of legal cases, audits, and appeals of systems.

In life-altering application domains, care must be taken to ensure that automated systems do not introduce systematic biases and harmful side effects. Medical devices and drugs undergo heavy regulation to prevent unverified claims; why should life-altering autonomous decision-making systems not bear a similar burden? People should not be guinea pigs for algorithms deployed untested in real-world settings, no matter how promising the application, or how well the technology has worked elsewhere. While AI systems should not be limited to a single scope, nor prevented from being applied outside of their originally-intended context, a system's efficacy ought to be re-validated in each new deployment context.

Requiring a validated scope for autonomous decision-making systems would place the burden of proof of their system's efficacy on developers wishing to enter the market, and on clients wishing to transfer an existing system to a new market, rather than on wronged subjects trying to prove out-of-scope use. Given the power that decision-making systems can have over livelihoods, providing evidence that the system is functioning as intended seems obvious, yet stories continue to emerge of developers and clients rushing to deploy biased or largely-untested decision-making systems. ProPublica's investigation into Northpointe's COMPAS algorithm found that it was only about 20% accurate at identifying violent reoffenders; the figure for non-violent reoffenders was just 61%. Worse, the system was twice as likely to rule unfavourably for black offenders than white offenders (Angwin & Larson, 2016).

Additionally, some recent systems are deployed based on spurious, pseudoscientific claims, such as Predictim's rating potential babysitters for "disrespectful attitude" (Merchant, 2018), or companies claiming to determine personality traits and even "criminality" from facial features (Storm, 2016) (y Arcas, Mitchell, & Todorov, 2017). Requiring systems to validate the scope of their application before deployment causes such premises to fall apart. Attempting to define a metric by which to evaluate disrespectfulness may reveal inherent biases, which the AI will subsequently learn. Such applications, which are designed to prey on fear rather than provide a truly beneficial service, will be forced to either move towards evidence-based validation and metrics, or to explicitly state that their system is an elaborate placebo with no real predictive power, which will lessen their appeal.

Evaluating the use cases under which an AI system performs well will also bring to light its limitations. Though popular culture pushes the narrative that AI is rapidly approaching general intelligence, the current reality is that AI achieves high performance only at very narrow tasks, and does not yet generalize to other tasks. This leads to fragile, brittle systems, as demonstrated by the effectiveness of adversarial attacks (Szegedy et al., 2013).

Without a clearly-defined scope, clients without AI expertise may misunderstand how a developer's system is intended to be used, or may misinterpret the AI's decision. For example, the Northpointe COMPAS recidivism algorithm was originally designed to inform treatment, but, unbeknownst to the developer, was being used for sentencing: in Wisconsin, a judge overturned a plea deal after viewing a defendant's COMPAS risk score (Angwin & Larson, 2016). Whether due to fragility, unstated limitations, or the nature of the data, decision-making systems may work in one situation but not another, yet clients who fail to understand these issues may take systems which are fair, unbiased, and valid in one context, and unwittingly deploy them in another context which violates one of these conditions.

A clearly-defined scope can also help protect developers from charges of misuse of personal data. Regulation surrounding the use of these data are tightening, as seen with the European Union's (EU's) new General Data Protection Regulation (GDPR, 2016). A developer that collected data under a certain pretense to build a model with a specific purpose may be at least partially liable to violations of that pretense if a client uses the model for another purpose. However, if developers explicate the purpose of their system ahead of time, the liability for misapplication rests entirely with the clients.

Proposed Framework

To bring these requirements together in a way that places the onus on developers and clients to jointly create fair, unbiased, accountable systems, we propose a formal regulation system that requires a pair of documents to be filed for any autonomous decision-making system to be deployed in a high-impact application domain.

The first document, referred to as an AI Validation Document (AIVD), is developer-filed and concerns transparency and validated scope: it defines one or more contexts for which the system was developed, outlines the claims made with respect to the system in each context, demonstrate how these claims were validated, and explains how the system's decisions can be interpreted in the case of an audit. The second document, referred to as a

Deployment Disclosure Document (DDD), is client-filed and concerns disclosure and auditability: it identifies the process for deploying a developer's AI in validated contexts and the specific terms surrounding disclosure and appeal of the AI's decisions. If this deployment is legally challenged (e.g. through the appeal of a decision that is alleged to be biased or unfair), these documents can help determine fault, if any. Though they would be filed separately, requiring both documents incentivizes developers and clients to communicate carefully regarding the accountability surrounding any particular deployment, and to understand both the technical workings of the AI, as well as the domain-specific policies and challenges related to the deployment context.

Specifically, an AIVD consists of defining:

- the intended purpose of the system, and
- the conditions under which it can be safely used for the intended purpose
- how those use cases have been tested and what metrics have been used to validate them, including how the results have been checked for bias
- known limitations of those use cases
- a description of the data used to train the system, including when, where, and how it has been sourced
- measures undertaken to track the model's development, including all versions of the model that may be deployed, and the specifics of how the data was used to train and validate each version
- how results of the system may be interpreted

This avoids problems such as Northpointe's COMPAS having an unvetted bias against black offenders, as such bias would need to be identified in COMPAS's AIVD, which would result in an undeployable system.

This document would have to be filed first, to establish the validity of the application in at least one context. In addition, an AIVD could be amended over time to accommodate new use cases whenever the developer can provide sufficient evidence supporting an extension of scope.

Once an AIVD has been approved, clients wishing to use one or more developer applications as part of a larger process or system would need to file a DDD containing:

- the purpose of the decision-making process to be created
- the reference number of the AIVD(s) being deployed
- the specific context under which each referenced system will be deployed within the larger application
- the process for disclosing the use of AI to subjects of the decision-making process
- the process by which a subject can investigate or appeal a decision
- the process by which the system's decisions will be traced and linked to the individual models involved in the decision-making process, including the assignment of responsibility between interacting models, or models making recommendations to a human reviewer

This avoids problems such as Northpointe's COMPAS being used by courts in sentencing, which was never intended by Northpointe and thus would not have been in the COMPAS's AIVD. Then, any courts trying to misapply COMPAS would have their DDD application denied.

One AIVD could be associated with many different DDDs, across a wide variety of countries and contexts, provided the processes described in the DDD fall within the validated scope outlined in the AIVD. This incentivizes developers to create large, reliable, and robust AI systems, since the marginal cost of an additional deployment is small compared to the initial cost of developing the system. This incentive also suggests that an international treaty or body setting the standards for evaluating these systems and their deployment is favourable, as it affords economic benefits to filing in member nations, akin to the system of international agreements surrounding intellectual property.

However, international cooperation is not integral to the framework. Instead, the AIVD and DDD can be filed on a national basis, with a national (or regional, as in the EU) body of experts reviewing AIVD and DDD applications. Approving AIVDs requires significant expertise in various domains, including AI, and AIVDs will be subject to more scrutiny and longer evaluation times. In contrast, smaller groups of legal and domain experts are sufficient to evaluate DDDs in a timely fashion.

In the spirit of promoting transparency and accountability throughout, both documents should contain highly-technical, legal sections for expert review and simple, nontechnical sections for subject review. In other words, the documents themselves must be transparent. This helps prevent the common scenario in which users blindly agree to certain documents, such software end-user license agreements, because they are too complicated and lengthy to be read and understood.

Concerns surrounding the stifling of innovation are apparent with the introduction of these documents. As such, we propose that decision-making systems deployed in lower- or questionable-impact domains would not be required to file these documents, though developers and clients using AI in these areas would still be incentivized to file for the additional legal protection. This raises an issue regarding the definition of the impact level of a system. After all, seemingly benign systems, such as recommender systems on social media, can potentially have a huge impact on the general public's views and perception of reality (Mozur, 2018), even though this impact may be difficult to quantify, especially on the individual level. Still, we believe requiring AIVDs and DDDs only for high-impact domains achieves the best tradeoff between innovation and responsible development. In particular, it helps mitigate any potential stifling of innovation in the startup sphere—where a small team may not have the necessary data accessibility or legal expertise required to file AIVDs and DDDs—unless the startup is innovating in a clearly high-impact domain, in which case provisions for dealing with accountability ought to factor into their business model from the start. AI researchers would also be exempt from filing these documents, regardless of the impact level of their research, as the use of AI in this context is accounted for through the proper informed consent of research subjects, as well as the research ethics approval process.

Outside of a research setting, certain clients may need to engage in pilot projects before formally deploying a decision-making system. In this case, a precursor to a DDD is to be filed, which describes the attempted purpose, a timeline for the completion of the pilot phase of the project (after which a full DDD must be filed), and the process for disclosure. Disclosure in pilot projects must additionally include provisions for soliciting and incorporating feedback and concerns from subjects. This ensures that experimental or fringe ideas involve a significant amount of shared decision-making between clients and subjects.

Benefits of a Two-Phase Framework

We argue that this proposal encourages developers to share knowledge with clients, and to engage with the consequences of their systems, while also ensuring clients have a framework with which to evaluate and question developers regarding any planned deployment. AI expertise is scarce compared to demand (Perry, 2018), and many clients outside of the information technology domain—which may include institutions such as hospitals, justice systems, and civic organizations—may not be able to acquire or retain such expertise. Regulation in the form of a DDD will push clients to choose vetted developers, who have filed AIVDs, instead of those making baseless claims. Hype around the power of big data and machine learning may engender blind trust in non-expert clients; this document system may raise awareness regarding the efficacy of AI systems, as clients must evaluate the AIVDs of developers to file their own DDD. Both clients and policymakers would be encouraged to see AI systems as procedures that may have flaws and should therefore be examined and challenged. As AI failures become more pronounced in the public eye, developers pushing AI without a AIVD, even in a lower-impact domain, may be seen as a liability, encouraging more thoughtfully-developed and carefully-tested AI.

Under this document system, profit-driven developers have an incentive to help non-expert clients define their disclosure, tracing and appeal processes for each of their AIVDs, in order to quickly file DDDs for as many clients as possible. In addition, developers wanting to work with clients outside the scope of their AIVD would have to devise mechanisms for efficiently validating these new use cases, leading to much-needed innovation in this area, while simultaneously promoting accountability. This also incentivizes developers to concern themselves with the real-world implications of their systems on subjects. Essentially, both developers and clients are incentivized to proactively establish fair and accountable applications of AI systems, instead of considering it an afterthought until bad press comes to light.

Formally-defined accountability documents that are widely recognized also have the potential to raise public awareness around the technical, legal and ethical issues of decision-making AI. Laypeople, who will likely be the subjects of decision-making systems, need to be informed about how these systems may be behaving (or misbehaving) and about their rights to appeal automated decisions. Much in the same way that common knowledge of copyright and patents brings awareness to issues of intellectual property, AIVDs and DDDs being commonly recognized as a crucial step of AI deployment will help subjects be aware of the requirement that these systems disclose the use of AI, are transparent, and are validated. This, in turn, will allow subjects to openly question institutions deploying AI, to be more involved in the development of automated processes, and to hold these institutions accountable.

Conclusion

As decision-making AI systems are becoming ubiquitous, concerns surrounding bias, fairness, and accountability are mounting. Transparency in these systems is critical in reducing their potential harm. However, transparency alone is not actionable without additional requirements to close the gap between developers, clients, policymakers, and the public, since developing and deploying fair and accountable AI systems requires increased awareness of the strengths and limitations of automated decision-making. For instance, clients may be unaware of the limitations of these systems, or of what constitutes fair and ethical AI. In contrast, developers may not understand how to create effective policy surrounding their systems' use, or how to deploy them safely in novel and untested contexts.

Introducing disclosure and scope into the AI system's development and deployment process not only encourages proactive collaboration between both parties, but also ensures subjects are made aware, and can appeal the decisions, of the system. Unlike transparency, where may open problems remain, disclosure and scope are primarily policy-based, and can thus be feasibly implemented in all AI systems. These requirements serve not only to render transparency algorithms actionable once properly developed, but also to educate all stakeholders about the potential and the pitfalls surrounding AI systems.

We proposed a regulatory framework for AI systems that formalizes the disclosure and scope requirements in the form of two documents, AIVDs and DDDs. The double-document structure of this framework incentivizes a more collaborative AI ecosystem, as well as ensuring that fairness and validity are not afterthoughts in high-impact AI systems. Fostering this dialogue between all stakeholders in a decision-making process spurs innovation, and will help developers, clients, policymakers, and the public realize the potential that effective, fair, and accountable automated systems have.

References

Angwin, J., & Larson, J. (2016, May). *Machine Bias.* Retrieved 2019-01-09, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Diallo, I. (2018, June). *The Machine Fired Me.* Retrieved 2019-01-02, from https://idiallo.com/blog/when-a-machine-fired-me

Freedland, J. (2018, April). Zuckerberg got off lightly. Why are politicians so bad at asking questions? *The Guardian*. Retrieved 2019-02-16, from https://www.theguardian.com/commentisfree/2018/apr/11/mark-zuckerberg-facebook-congress-senate

GDPR. (2016, May). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec. *Official Journal of the European Union*, *L119*, 1–88.

Good, B. J. (1993). *Medicine, rationality and experience: an anthropological perspective*. Cambridge University Press.

Kang, C., Kaplan, T., & Fandos, N. (2018, April). Knowledge Gap Hinders Ability of Congress to Regulate Silicon Valley. *The New York Times*. Retrieved 2019-02-16, from https://www.nytimes.com/2018/04/12/business/congress-facebook-regulation.html

Merchant, B. (2018, December). *Predictim Claims Its AI Can Flag 'Risky' Babysitters. So I Tried It on the People Who Watch My Kids.* Retrieved 2018-12-27, from https://gizmodo.com/predictim-claims-its-ai-can-flag-risky-babysitters-so-1830913997

Miller, T., Howe, P., & Sonenberg, L. (2017, December). Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv:1712.00547 [cs]*. Retrieved 2019-01-08, from http://arxiv.org/abs/1712.00547 (arXiv: 1712.00547)

Mozur, P. (2018, October). A Genocide Incited on Facebook, With Posts From Myanmar's Military. *The New York Times*. Retrieved 2019-01-11, from https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

Perry, T. S. (2018, September). *Intel Execs Address the AI Talent Shortage, AI Education, and the "Cool" Factor.* Retrieved 2019-01-10, from https://spectrum.ieee.org/view-from-the-valley/robotics/artificial-intelligence/intel-execs-address-the-ai-talent-shortage-ai-education-and-the-cool-factor

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (pp. 1135–1144). San Francisco, California, USA: ACM Press. Retrieved 2019-01-08, from http://dl.acm.org/citation.cfm?doid=2939672.2939778 doi: 10.1145/2939672.2939778

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in neural information processing systems* (pp. 2503–2511).

Smith, A. (2018, November). *Public Attitudes Toward Computer Algorithms.* Retrieved 2019-01-05, from http://www.pewinternet.org/2018/11/16/attitudes-toward-algorithmic-decision-making/

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008, May). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*(5), 543–545. Retrieved 2019-02-16, from http://www.nature.com/articles/nn.2112 doi: 10.1038/nn.2112

Storm, D. (2016, May). *Faception can allegedly tell if you're a terrorist just by analyzing your face.* Retrieved 2019-01-01, from https://www.computerworld.com/article/3075339/security/faception-can-allegedly-tell-if-youre-a-terrorist-just-by-analyzing-your-face.html

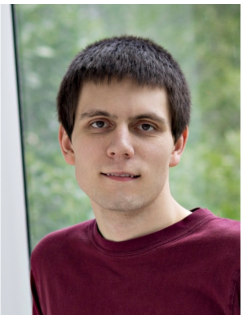
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv:1312.6199 [cs]*. Retrieved 2019-01-08, from http://arxiv.org/abs/1312.6199 (arXiv: 1312.6199)

Tversky, A., & Kahneman, D. (1981, January). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. Retrieved 2019-02-16, from http://www.sciencemag.org/cgi/doi/10.1126/science.7455683 doi: 10.1126/science.7455683

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Mathur, V., West, S. M., ... Schwartz, O. (2018, December). *AI Now Report 2018* (Tech. Rep.). AI Now Institute, New York University. Retrieved from https://ainowinstitute.org/AI_Now_2018_Report.pdf

y Arcas, B. A., Mitchell, M., & Todorov, A. (2017, May). *Physiognomy's New Clothes.* Retrieved 2019-01-07, from https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

Janelle Berscheid is a Master's student in the Computational Epidemiology and Public Health Informatics Lab (CEPHIL) at the University of Saskatchewan. Her current research focuses on health-related applications of data science and machine learning, and how these technologies can impact healthcare delivery.



Francois Roewer-Despres is a Master's student at the University of Toronto and a member of the Vector Institute for Artificial Intelligence. His research interests broadly cover Human-AI Interaction, including accountability and safety, autonomous social intelligence, dialogue systems, and reinforcement learning.

Context-conscious fairness in using machine learning to make decisions

Michelle Seng Ah Lee (University of Oxford; michelle.lee@oii.ox.ac.uk)

DOI: 10.1145/3340470.3340477

Abstract

The increasing adoption of machine learning to inform decisions in employment, pricing, and criminal justice has raised concerns that algorithms may perpetuate historical and societal discrimination. Academics have responded by introducing numerous definitions of “fairness” with corresponding mathematical formalisations, proposed as one-size-fits-all, universal conditions. This paper will explore three of the definitions and demonstrate their embedded ethical values and contextual limitations, using credit risk evaluation as an example use case. I will propose a new approach - context-conscious fairness - that takes into account two main trade-offs: between aggregate benefit and inequity and between accuracy and interpretability. Fairness is not a notion with absolute and binary measurement; the target outcomes and their trade-offs must be specified with respect to the relevant domain context.

Introduction

Machine learning (ML) is increasingly used to make important decisions, from hiring to sentencing, from insurance pricing to providing loans. There has been an explosion of publications of new guidelines and principles of Artificial Intelligence (AI) from governments (Holden & Smith, 2016), international organisations (European Commission AI HLEG, 2019), professional associations (IEEE Global Initiative, 2018), and academic institutions (Grosz, 2015). Each of these documents references the imperative for AI to treat people fairly by avoiding discrimination based on legally protected characteristics, such as race, gender, and sexual orientation. However, none of them specifies a framework or methodology to detect and correct unfair outcomes, especially given history of discrimination in our systems and societies that predates the introduction of algorithmic decision-making. Given a bias, people-based processes may arrive at different decisions. AI, by contrast, can replicate an identical bias at-scale, crystallising the bias and removing the outcome ambiguity associated with human decision-making. This is especially concerning in domain areas with documented historical discrimination, as AI can exacerbate any underlying societal problems and inequalities. For example, discriminatory lending has been a contentious problem. In the United States, ‘redlining,’ risk inflation of minority-occupied neighbourhoods, impeded African Americans from obtaining mortgage (Nelson, Winling, Marciano, & Connolly, 2016). Analysis of 2001-2009 UK consumer credit data on 58,642 households found that non-white households are less likely to have financing (Deku, Kara, & Molyneux, 2016). To counteract this, new techniques have been introduced for *pre-processing* (purging the data of bias prior to training the algorithm) and for *post-processing* (bias correction in predictions after algorithm build). In both cases, solutions are based on a one-size-fits-all condition of *fairness*, regardless of the context.

There is an overall consensus that AI systems and technologies should be required to make fair decisions, yet the binary categorisation of an algorithm as either fair or unfair belies the underlying complexities of each use case. In the next section, I will use discriminatory lending as an example to demonstrate the shortcomings in existing attempts at formalising fairness in machine learning predictions. Different issues may rise in other domains, such as employment, pricing, or criminal justice; however, the focus on one use case will help bring to light the real-life considerations of each fairness definition. In Section 3, I will propose a new approach focusing on quantifying the contextual trade-offs of each algorithm so that the decision-makers can select a model that best captures the risk profile of each specific case.

Ethics of fairness definitions and their limitations

If the outcome disparity in loan decisions is simply a factor of significant features (e.g., difference in income), arguably the variance in outcome may be consistent with the underlying distributions. Given that minority borrowers are indeed riskier, from an expected consequentialist perspective, overall societal welfare would increase with a more accurate prediction of default risk. If we subscribe to this, the machine learning model is fair insofar as it most accurately predicts risk and gives each individual the loan and the interest rate that he or she deserves, by foreseeing the consequences of a loan approval. However, the outcome disparity could be a reflection of inequality in other markets (e.g. in labour markets), which makes minority incomes more volatile. Given the evidence in past literature of historical discrimination based on race and gender in mortgage lending decisions in the US (Kendig, 1973), a machine learning algorithm can self-perpetuate a reproduction of past inequalities, inaccurately inflating the risk of minority borrowers. The extent of this inaccuracy is difficult to determine; it is impossible to know whether those who were denied a loan would have defaulted. The scarcity and potential bias of historical data on previously financially marginalised groups hinders the statistical modelling of their credit-worthiness. If historical evaluations of minority credit risk are inaccurate, a machine learning model trained on a biased data set necessarily produces a sub-optimal result. I will discuss three of the main approaches that seek to define whether or not an outcome is unfair.

Demographic parity

A strictly egalitarian approach mandates *equal outcome* for each racial group. A related statistical definition is *demographic parity*, a population-level metric that requires the outcome to be independent of the protected attribute. Formally, with Ŷ as the predicted outcome and A as a binary protected attribute, we have:

$$P(\hat{Y}|A=0) = P(\hat{Y}|A=1) \tag{1}$$

While this metric would ensure the equally proportional outcome independent of race, it is ineffective where disproportionality in outcomes can be justified by non-protected, non-proxy attributes, as this can lead to reverse discrimination and inaccurate predictions (Gajane, 2017). Fuster et al. have shown that there is a difference in income distributions between racial groups (Fuster, Goldsmith-Pinkham, Ramadorai, & Walther, 2017). As income has a reasonably inferential relationship to credit risk, it cannot be considered as a simple proxy attribute of protected characteristics. In short, implementing this measure would unfairly give credit to minorities with low income.

Equalised opportunity

According to a Rawlsian approach to distributive justice, regardless of whether policymakers have the moral imperative to correct past wrongs of systematic discrimination, it is important to prioritise the protection of the most vulnerable of the population. The Max-Min social welfare function maximises the welfare of those who are worst-off (Rawls, 1971). In contrast to equal outcome, Rawls' Difference Principle asserts the need for *equality in opportunity*.

Hardt, Sbrero and Price proposed a statistical metric of "equalised opportunity" that focuses on the true positives: given a positive outcome, the prediction is independent of the protected attribute. Formally, again with Ŷ as the predicted outcome, A as a binary protected attribute, and Y as the actual outcome, we have:

$$\begin{aligned} &P(\hat{Y}=1|A=0,Y=1) \\ &= P(\hat{Y}=1|A=1,Y=1) \end{aligned} \tag{2}$$

Unfortunately, equalised opportunity metric is also problematic because it fails to address discrimination that may already be embedded in the data (Gajane, 2017). Gender and race are outside of one's control, and following the logic of Dworkin's theory of Resource Egalitarianism, no one should end up worse off due to bad luck, but rather, people should be given differentiated economic benefits as a result of their own choices (Dworkin, 1981). The credit risk market does not exist in a vacuum; while people can affect their credit scores and income to a certain extent, e.g. by building their credit history or improving employable skills, it is impossible to isolate similar variables from the impact of their upbringing, discrimination in other markets, and historical inequalities entrenched in the data.

Counterfactual fairness

A challenge to the previous two fairness metrics is that they do not take seriously the notion of causality. Statistical relationships are derived from correlations rather than an inferential and directed model. David Lewis (1973) introduced the counterfactual approach of theorising on the cause and effect (Lewis, 1973). *Counterfactual fairness*, one of the most recently introduced definitions, attempts to apply this theory to assert that a decision is fair if it is the same in the actual world as it would be in a counterfactual world where the individual belonged to a different demographic group (Kusner, Loftus, Russell, & Silva, 2017). Given a causal model with latent background variables U, non-proxy non-protected features X, and the protected attribute A, Ŷ as the predicted outcome, and Y as the actual outcome, we have:

$$\begin{aligned} &P(\hat{Y}_{A\leftarrow a}(U) = Y|X = x, A = a) \qquad (3)\\ &= P(\hat{Y}_{A\leftarrow a'}(U) = y|X = x, A = a)\end{aligned}$$

However, this approach requires a formalisation of causal models, which are useful abstractions but often infeasible without strong assumptions, simplifications, and robust constraints. In reality, relationships between variables are intertwined, complex, or not fully understood. For example, how would we isolate the impact of one's race on credit-worthiness from the impact of one's education, job, neighbourhood, and other features of one's upbringing? There may also be challenges of reverse causality: e.g., race may impact one of the non-protected predictors, e.g. if employers are discriminatory, so that an applicant's low income ends up hindering both the probability of loan approval and his or her ability to repay a loan.

Implementing a "fair" credit risk algorithm is difficult not just because there are competing theoretical approaches to fairness, but because they all assume that there is a one-size-fits-all definition of fairness. What seems to be required is a new approach that builds on the previous ones but makes the most of the context-dependency of the data available and hence of the relational nature of fairness (*a*) is fair - not absolutely - but in relation to (*b*). I introduce a possible proposal in the next section.

Proposed solutions

While the implications of unequal outcome in mortgage lending are troubling, the revelation is also an opportunity for policymakers to redefine anti-discrimination policies. Consider the alternatives to a machine learning model. Human-driven decision-making is mired in cognitive biases, as shown by a field experiment on discrimination in the labour market, in which white-sounding names received 50% more callback than black-sounding names (Bertrand & Mullainathan, 2003). Traditional credit risk models, e.g. a scorecard, are also not exempt from selection and counterfactual biases of a machine-learning model. The accusation of discrimination in mortgage lending long predates the introduction of machine learning; bias can be embedded in a process (Kendig, 1973).

In reality, the alternative to a machine learning model may be a worse model. Therefore, I propose a benchmarking exercise to map the change in the two trade-offs, while increasing the model complexity. While no model, pre-processed or post-processed, may eliminate all the discriminatory impact on minority groups, quantifying the benefits and risks would give actionable insights to the decision-maker on which model best reflects his or her values and risk profile. Algorithmic processing of credit worthiness would not translate into automatic decision making but would empower a better evaluation of each case under scrutiny.

The shift of focus from discriminatory intent to discriminatory impact in scholarship is mirrored in legal rulings. "Fairness through Unawareness" approach - not using the protected characteristic - is invalidated with ML models that can triangulate protected information.[1] Even if a lender is not explicitly considering race in calculating credit risk, a machine learning model may nonetheless incorporate

[1] See: (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2011) for an extensive critique.

racial information in its prediction. Given legal decisions in the UK (Lowenthal, 2017) and in the US (Baum & Stute, 2015), discriminatory impact is illegal regardless of intent. Both courts emphasise a contextual exception: disparate impact is allowed if it is crucial to a legitimate business requirement. One of such requirements may be the need to predict accurately default to allow for greater financial inclusion, which - given discriminatory history - may be at the expense of some minority groups.

Trade-off 1: Aggregate benefit vs. individual inequity

Credit market welfare can be measured by the consumer surplus generated. Automated underwriting improves the overall accuracy of default predictions (Hacker & Wiedemann, 2017). This increases the market information available to the lenders and lowers their overall risk, empowering them to give credit to those who were previously unable to access it. The resulting increase in financial inclusion adds to the overall consumer surplus.

On an individual level, however, there are winners and losers: the borrowers who would not have received a loan under traditional technology who now successfully secure one; and the potential borrowers who would have received a loan before but are deemed too high risk under the new technology. A troubling finding of Fuster et al. is that the losers are disproportionately from minority racial groups, likely as a reflection of historical discrimination that is inaccurately inflating the risk of minority borrowers (Fuster et al., 2017). Meanwhile, the various attempts at an active correction of this "unfairness", synthesised in Section 2, have only obfuscated the criteria for fairness, as they do not provide a satisfactory framework to determine which definition is most appropriate for a given use case. Mapping of this trade-off between aggregate benefit (e.g. increased financial inclusion) and inequity (e.g. negative impact on minority borrowers) would allow the lender to identify the most appropriate algorithm to model default risk.

Figure 1 is an indicative visualisation of this trade-off. There are two assumptions that should be tested: 1) aggregate benefit improves with the complexity of the model and 2) The increase in benefit is at the expense of the welfare of the protected group. The first assumption is reasonable for most complex models in which an implementation of a machine learning algorithm is being considered. If the true relationship is linear, then the out-of-sample accuracy would decrease with additional complexity, but this will be rare in many real-life modelling scenarios. The second assumption would depend on the joint distribution of the protected characteristic and the outcome being predicted. In many cases, the risk of discrimination exists because of the statistical disparity in outcomes between groups.

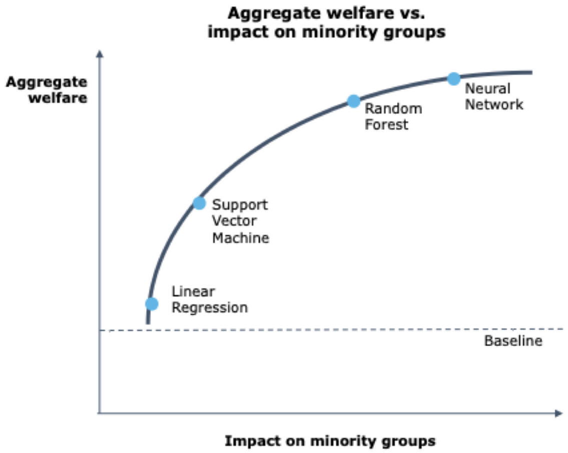


Figure 1: Benchmarking models for trade-off in aggregate group benefit vs. the scale of negative impact on protected groups

Rather than enforcing a single definition of fairness, this allows a context-conscious analysis of which model best serves the customers or society. This exercise should give a decision-maker an insight into the model and the trade-off with which he or she is the most comfortable. Once these trade-offs are recognised and quantified, it is important to consider whether the inter-group outcome disparity is fair and justified.

Trade-off 2: Accuracy vs. interpretability

Historically, lenders have focused on data that theoretically relate to outcome; for example, the debt-to-income ratio and past payments are indicators of default risk. With Big Data analytics, firms are beginning to incorporate non-traditional data types into their algorithms as proxies of risk. An extreme case is the use

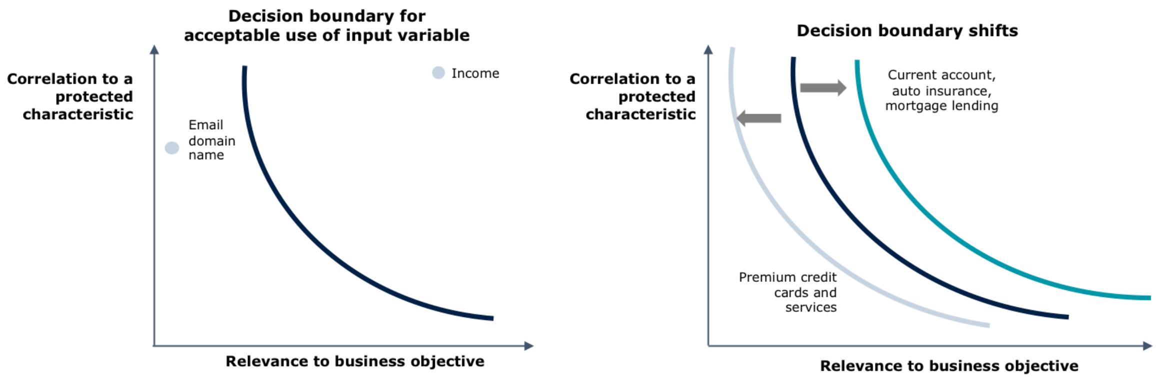

Decision boundary for acceptable use of input variable
Correlation to a protected characteristic
Income
Email domain name
Relevance to business objective
Decision boundary shifts
Correlation to a protected characteristic
Current account, auto insurance, mortgage lending
Premium credit cards and services
Relevance to business objective

Figure 2: Decision boundary for acceptable use of feature

of Chinese citizens' Internet browsing history, location, and payment data to calculate credit risk (Koren, 2016). While the justification is to open up access to credit to those without credit history, this data use risks unfairly marginalising people based on their background, in uninterpretable and opaque ways. Previously, a prospective borrower could build his or her credit history by paying bills on time and expect a positive impact on the probability of loan approval. With Big Data and machine learning, this expectation may be unfounded.

A decision-maker must justify whether the outcome disparity is due to legitimate and non-discriminatory difference in the outcome distributions. Figure 2 visualizes a possible decision boundary for whether or not an input variable should be used in a model, based on its role as a potential proxy for a protected characteristic. Given Supreme Court decisions in the UK (Lowenthal, 2017) and the US (Baum & Stute, 2015), even if a variable is correlated to a protected feature, there may be reasonable grounds to use it if the differences are crucial to a legitimate business requirement. This decision boundary may shift depending on the context. The drivers of decision-making in providing essential products, such as current account, car insurance, or mortgage, may be subject to higher scrutiny than the rationale for offering premium credit cards. This ensures that the decision to include features correlated to protected characteristic is carefully considered within the context of the regulated domain and the potential impact on the customers.

Policymakers should consider this trade-off between accuracy and interpretability to limit what data can be used and what models can be built. For provision of essential products, e.g. current accounts, criminal justice decisions, and car insurance, that have a significant impact on people's lives, policymakers may need to ensure that all features included in the variable have a strong inferential relationship with the outcome rather than simply for predictive correlation.

Conclusion

The introduction of ML into lending models presents a new opportunity for greater accuracy in predicting default and, therefore, greater financial inclusion. However, the prevalence of ML models in crucial decision-making processes has brought to light the intricate ways in which discrimination can be perpetuated through these technologies. Fairness is a complex feature of the world, and no single definition can mitigate the risk of unfair treatment.

At the same time, ML models are auditable for fairness. The opportunity lies in the computational and systematic decision-making of ML. A more context-conscious approach of benchmarking the trade-offs between aggregate benefit and individual inequity and between accuracy and interpretability would force the decision-maker to identify the model that is most suitable to each case. The risk of discrimination in ML is undeniable; however, it presents an opportunity to consider more closely what we value as a society and to pursue fair treatments and decisions that can be enforced in our markets.

References

Baum, J. S. J. J., D., & Stute, D. (2015). Supreme court affirms fha disparate impact claims.

Bertrand, M., & Mullainathan, S. (2003, July). *Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination* (Working Paper No. 9873). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w9873 doi: 10.3386/w9873

Deku, S. Y., Kara, A., & Molyneux, P. (2016). Access to consumer credit in the uk. *The European Journal of Finance*, *22*(10), 941-964. Retrieved from https://doi.org/10.1080/1351847X.2015.1019641 doi: 10.1080/1351847X.2015.1019641

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2011). Fairness through awareness. *CoRR*, *abs/1104.3913*. Retrieved from http://arxiv.org/abs/1104.3913

Dworkin, R. (1981). What is equality? part 1: Equality of welfare. *Philosophy and Public Affairs*, *10*(3), 185–246.

European Commission AI HLEG. (2019). *Draft ai ethics guidelines for trustworthy ai* (Tech. Rep.). European Commission High-Level Expert Group on Artificial Intelligence (AI HLEG). Retrieved from https://ec.europa.eu/futurium/en/node/6044

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2017, 01). Predictably unequal? the effects of machine learning on credit markets. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3072038

Gajane, P. (2017). On formalizing fairness in prediction with machine learning. *CoRR*, *abs/1710.03184*. Retrieved from http://arxiv.org/abs/1710.03184

Grosz, R. A. E. H. A. M. T. M. D. M. Y. S., Barbara. (2015). *Artificial intelligence and life in 2030* (Tech. Rep.). Stanford University. Retrieved from https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fnl.pdf

Hacker, P., & Wiedemann, E. (2017). A continuous framework for fairness. *CoRR*, *abs/1712.07924*. Retrieved from http://arxiv.org/abs/1712.07924

Holden, J., & Smith, M. (2016). *Preparing for the future of ai* (Tech. Rep.). U.S. Executive Office of the President, National Science and Technology Council Committee on Technology. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

IEEE Global Initiative. (2018). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems.* Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

Kendig, D. (1973). Discrimination against women in home mortgage financing. *Yale Review of Law and Social Action*, *3*(2).

Koren, J. R. (2016, Jul). What does that web search say about your credit? Retrieved from https://www.latimes.com/business/la-fi-zestfinance-baidu-20160715-snap-story.html

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017, March). Counterfactual Fairness. *arXiv e-prints*, arXiv:1703.06856.

Lewis, D. (1973). Causation. *Journal of Philosophy*, *70*(17), 556–567.

Lowenthal, T. (2017). Essop v home office: Proving indirect discrimination.

Nelson, R. K., Winling, L., Marciano, R., & Connolly, N. (2016). Mapping inequality. *American Panorama*. Retrieved from https://dsl.richmond.edu/panorama/redlining/#loc=5/36.721/-96.943&opacity=0.8&text=intro

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Michelle Seng Ah Lee
michelle.lee@oii.ox.ac.uk

Michelle is an MSc candidate in Social Data Science at the Oxford Internet Institute and a Research Assistant at the Digital Ethics Lab, supervised by Professor Luciano Floridi. Her research focuses on fairness in machine learning algorithms and their trade-offs on aggregate and individual levels. She is also interested in the regulatory and ethical considerations of bias and discrimination in artificial intelligence (AI).

The Scales of (Algorithmic) Justice: Tradeoffs and Remedies

Matthew Sun (Stanford University; mattsun@stanford.edu)
Marissa Gerchick (Stanford University; gerchick@stanford.edu)

DOI: 10.1145/3340470.3340478

Introduction

Every day, governmental and federally funded agencies — including criminal courts, welfare agencies, and educational institutions — make decisions about resource allocation using automated decision-making tools (Lecher, 2018; Fishel, Flack, & DeMatteo, 2018). Important factors surrounding the use of these tools are embedded both in their design and in the policies and practices of the various agencies that implement them. As the use of such tools is becoming more common, a number of questions have arisen about whether using these tools is fair, or in some cases, even legal (*K.W. v. Armstrong*, 2015; ACLU, Outten & Golden LLP, and the Communications Workers of America, 2019).

In this paper, we explore the viability of potential legal challenges to the use of algorithmic decision-making tools by the government or federally funded agencies. First, we explore the use of risk assessments at the pre-trial stage in the American criminal justice system through the lens of equal protection law. Next, we explore the various requirements to mount a valid discrimination claim — and the ways in which the use of an algorithm might complicate those requirements — under Title VI of the Civil Rights Act of 1964. Finally, we suggest the adoption of policies and guidelines that may help these governmental and federally funded agencies mitigate the legal (and related social) concerns associated with using algorithms to aid decision-making. These policies draw on recent lawsuits relating to algorithms and policies enacted in the EU by the General Data Protection Regulation (GDPR) (2016).

Algorithms and Equal Protection

One case of algorithmic decision-making in the public domain that has been recently subjected to increased scrutiny in recent years is the use of risk assessments in the criminal justice system. Here, we focus on the use of criminal risk assessment at the pre-trial stage. The goal of risk assessment tools (RATs) at the pre-trial stage is typically to estimate a defendant's likelihood of engaging in a particular future action (for example, committing a new crime or failing to appear in court) based on their similarity to defendants who have committed those actions in the past (Summers & Willis, 2010). This similarity is typically determined using factors regarding a defendant's criminal history but may also include information about a defendant's personal and social history such as their age, housing and employment status, and in some cases, their gender (Summers & Willis, 2010; *State v. Loomis*, 2016). Risk assessments are not themselves decision-makers regarding detention; rather, they are tools used by a human decision-maker - typically a judge or magistrate (Desmarais & Lowder, 2019).

In this section, we explore legal challenges pertaining to risk assessments on the basis that their use, under some circumstances, may violate constitutional protections. In particular, the Fifth Amendment guarantees equal protection under due process of law and applies to the federal government (U.S. Const., amend. V.), while the Fourteenth Amendment guarantees equal protection and due process of law and applies to the states (U.S. Const., amend. XIV.). Our analysis focuses on the application of equal protection law to the use of algorithmic risk assessments. Specifically, we discuss policies around the use of gender and proxies for race in risk assessments and how each might interact with equal protection of the law.

When an individual or entity believes that their right to equal protection has been violated by a governmental policy - such as the use of a risk assessment algorithm at the pretrial stage - they may challenge such a policy by, first, proving that the policy does indeed discriminate in a way that is or was harmful to the individual (Legal Information Institute, 2018a).The court evaluating the matter would then analyze the policy in question through one of four possible lenses - strict scrutiny, intermediate scrutiny, rational basis scrutiny, or a combination of the prior three, depending on the characteristic (race, national origin, gender, etc.) in question (Legal Information Institute, 2018a).

One such notable challenge, which we reference in the subsequent discussion, was *State of Wisconsin v. Loomis* (2016), in which Eric Loomis challenged the use of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment to inform a judge's decision about how long his prison sentence would be. Loomis challenged the use of COMPAS on the grounds that it violated his constitutional right to due process because the tool itself was proprietary (in particular, Loomis knew the factors used on the assessment but did not know how each of those factors was weighted and translated into a score, and thus could not challenge its scientific validity), and because the tool used gender as a factor in the assessment (*State v. Loomis*, 2016).

Factor 1: Use of gender

Though many risk assessments used at the pretrial stage in the United States do not include gender as a factor in the calculation of risk scores (Latessa, Smith, Lemke, Makarios, & Lowenkamp, 2009; VanNostrand et al., 2009), some pretrial risk assessments do consider gender, like COMPAS did in the case of Eric Loomis (*State v. Loomis*, 2016). Moreover, evidence indicates that risk assessments may not be equally predictive across genders, and may overestimate the recidivism risk of women compared to men (Skeem, Monahan, & Lowenkamp, 2016). Such evidence suggests the counterintuitive idea that including gender in the calculation of risk scores may be more equitable than excluding it. To illustrate the complexities of this point, we consider two hypothetical scenarios regarding risk assessments and gender.

Consider a hypothetical risk assessment X that includes gender in its calculation of risk scores; assume X has been challenged on the basis that its use of gender violates equal protection. Equal protection claims involving gender classifications are subject to intermediate scrutiny, a test established by the Supreme Court in *Craig v. Boren* (1976). To pass intermediate scrutiny, the policy in question must ”advance an important government interest” by means that are ”substantially related to that interest” (Legal Information Institute, 2018b; *Craig v. Boren*, 1976). The defendant (the jurisdiction that uses X to inform pretrial release decisions) might argue that, because judges rely on the accuracy of risk scores when making decisions about who to release and because these risk scores are meant to inform their decision-making, the use of gender in X advances an important government interest - ensuring public safety through release determinations. The defendant might also argue that, given the evidence on differential predictive power by gender, the use of gender is indeed a means that is ”substantially related” to public safety.

In the case of Loomis, the court determined the use of gender was permissible because it improved accuracy, a non-discriminatory purpose (*State v. Loomis*, 2016). Yet some argue that such evidence regarding the differential predictive power by gender is too general. Legal scholar Sonja Starr has argued that because the Supreme Court has rejected the use of broad statistical generalizations about groups to justify discriminatory classifications, the use of gender in risk assessment (specifically at sentencing) is unconstitutional (Starr, 2014). In the case of X, the court would have to consider, given the relevant evidence, if it is actually the case that using gender as a factor is substantially related to public safety, weighing the tension between the group classifications in X and the principle of individualized decision-making in the criminal justice system.

Now consider risk assessment Y, a risk assessment that doesn't include gender in its calculation of risk scores, and suppose that a jurisdiction that uses Y has analyzed its own data and found that Y is better at predicting recidivism for men than it is at predicting recidivism for women. In this case, the policy in question is facially neutral (the use of Y doesn't appear to be discriminatory towards women and doesn't specifically include gender in its calculations), but nonetheless has a disparate impact because it rates women as higher risk than they actually are. If the use of Y were challenged under equal protection, the challenger would have to show intent - in particular, that the governmental body using Y intended to discriminate against women by using Y. In *Personnel Administrator of Massachusetts v. Feeney* (1979), the Supreme Court was faced with the question of whether a facially neutral policy that had a disparate impact on women was a violation of equal protection. A key question was whether the ”foreseeability” of the policy's disparate impact was sufficient proof of discriminatory intent; the court held that it was not (Weinzweig, 1983). Thus, if the ruling from Feeney were applied to the hypothetical case regarding Y, awareness of Y's differential predictive power for men and women may not necessarily qualify as proof of intent to discriminate, and the equal protection claim against Y may fall short.

Factor 2: Use of proxies for race

Now consider a hypothetical risk assessment Z that uses factors such as the stability of a defendant's housing or their employment status - in practice, many risk assessments do consider these factors, as they are correlated with recidivism risk (Summers & Willis, 2010). However, these factors may serve as proxies for race (Barocas & Selbst, 2016; Corbett-Davies & Goel, 2018). Though classifications involving race or national origin are typically subject to strict scrutiny, absent an explicit discriminatory classification, both disparate impact and discriminatory intent are required to even trigger a scrutiny test (as they would be in the hypothetical case of Y, described above) (*Arlington Heights v. Metropolitan Housing Dev. Corp.*, 1977). Thus, for Z's use to be challenged because of its use of proxies for race, one would need to show both that Z has a disparate impact (for example, that though scores inform decision-making for all people, Z is less accurate for minorities than for white people, which may or may not be true in the case of this hypothetical) and that Z was designed or used to be discriminatory against the minority group(s) in question. Demonstrating this intent may prove challenging because of the correlation between these socioeconomic factors and recidivism risk; nonetheless, the tension between statistical generalizations about groups of people and the right to an individualized decision for each defendant is ever present.

More broadly, legal challenges to the use of RATs under constitutional law speak to an underlying theme of the use of algorithms more generally: the use of these tools does not fit neatly into established legal standards (Barocas & Selbst, 2016), and tradeoffs will be present, whether mathematical, social, both, or otherwise (Corbett-Davies & Goel, 2018). Moreover, in the presence of facially neutral RATs, understanding intent is crucial to understanding if the law has been violated. In the remedies section, we propose inquires around RAT implementation that may help clarify the intent of policymakers and agencies who adopt these tools and inform the public about the agencies' decision-making rationale in the presence of tradeoffs.

Algorithms and Civil Rights Law

Beyond the constitutional arena, disparate impact theory has another, distinct form in civil rights law. Famously, Title VII of the Civil Rights Act of 1964 explicitly bars employment practices that would generate a disparate impact, defined by the following conditions: 1) the policy creates an adverse effect that falls disproportionately upon a particular protected class, 2) the specific policy in place is not a ”business necessity,” and 3) there exists an alternative policy that would not result in disproportionate harms (42 U.S.C. §2000e et seq.). In *Griggs v. Duke Power Co.* (1971), the Supreme Court found that Duke Power's requirement of a high school diploma for its higher paid jobs was illegal under Title VII of the Civil Rights Act of 1964 because it disproportionately barred minority groups from those positions and did not have any demonstrable relation to performance on the job.

Beyond Title VII and employment practices, the Court has ruled in multiple cases involving federal statutes with disparate impact provisions, such as *Lau v. Nichols* (1974) and *Alexander v. Choate* (1985), that policies which create adverse disparate impact are in violation of the law, regardless of the intent of those policies or whether the policies are applied equally to all groups. Such policies that create a disparate impact constitute a violation of Title VI of the Civil Rights Act of 1964, which was enacted at the same time as Title VII (42 U. S. C. § 2000d). We choose to now shift focus to Title VI because Title VI stipulates that all programs or activities that receive federal funding may not perpetrate or perpetuate discrimination on the grounds of race, color, or national origin, while Title VII only concerns employment (42 U. S. C. § 2000d). However, we note that the U.S. Department of Justice has recently stated that Title VI ”follows...generally..the Title VII standard of proof for disparate impact”; thus, cases that concern Title VII ”may shed light on the Title VI analysis in a given situation” (U.S. Department of Justice, 2019).

Twenty-six federal agencies have Title VI regulations that address the disparate impact standard, including USDA, the Department of Health and Human Services, and the Department of Education (U.S. Department of Justice, 2019). These federal agencies provide funding to a massive array of public programs and the social safety net, including public schools, Medicaid, and Medicare. In *Lau v. Nichols* (1974), for example, the Court found that the San Francisco Unified School District was in violation of Title VI because it received federal funding yet imposed a disparate impact on non English-speaking students, many of whom were not offered supplemental language instruction or placed into special education classes.

This regulatory and legal landscape sets the stage for the application of disparate impact theory under civil rights law as an important possible remedy for discrimination in algorithmic decision-making. As state and local governments increasingly turn towards automated tools to lower costs, ease administrative burdens, and deliver benefits, we are likely to observe cases where algorithms, especially when deployed without comprehensive oversight and auditing processes in place, create unequal outcomes. In her book *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, Professor Virginia Eubanks examines a statistical tool used by the Allegheny County Office of Youth, Children, and Families that processes data from public programs to predict the likelihood that child abuse is taking place in individual households across the county (Eubanks, 2018). Because the frequency of calls previously made on a family is an input to the algorithm, Eubanks argues that the tool may systematically discriminate against Black families, since Black families are far more likely to be called on by mandatory reporters or anonymous callers (Misra, 2018). The Office of Youth, Children, and Families is overseen by the Allegheny County Department of Human Services, which receives federal funding and as a result may be subject to regulation under Title VI (Allegheny County, 2019).

In these cases and many others, there is often no obvious evidence of discriminatory intent; to the contrary, algorithms are commonly deployed in the hopes of mitigating human biases (Lewis, 2018). In Allegheny County, officials stressed that the predictive risk-modeling tool would guide, not replace, human decision-making (Hurley, 2018; Giammarise, 2017). Yet, we often see that algorithms may still produce significant adverse impact on populations when analyzed on the basis of race or gender. As a result, groups or individuals may naturally seek to challenge the use of such algorithms in programs receiving federal funding under Title VI. According to a Justice Department legal manual on Title VI, three conditions are required to constitute a violation of Title VI: 1) statistical evidence of disparate adverse impact on a race, color, or national origin group, 2) the lack of a substantial legitimate justification for the policy, and 3) the presence of a less discriminatory alternative that would achieve the same objective but with less of a discriminatory effect (42 U. S. C. § 2000d).

In the following sections, we explore how disparate impact claims against the usage of algorithms might fail to succeed in court for three separate reasons. These challenges can be summarized as the lack of presence of a less discriminatory alternative, the use of predictive accuracy as ”substantial legitimate justification” for the policy, and the possibility that the only way to ameliorate disparate impact would be to treat different groups differently, thus triggering a disparate treatment legal challenge. We explore the current standard for how a complainant (i.e., plaintiff) must prove disparate impact under Title VI, and how a recipient (i.e., defendant) might ultimately circumvent their claims.

Challenge 1: Proving the presence of a less discriminatory alternative

The phrase ”less discriminatory alternative” implies that there exists a way to compare a set of policies and determine which is the least discriminatory. However, when it comes to algorithmic decision-making, the definition of ”fairness” (in other words, the absence of discrimination) is hotly debated (Gajane & Pechenizkiy, 2017). For example, the notion of ”classification parity” is defined as the requirement that certain measures of predictive performance, such as the false positive rate, precision, and proportion of decisions that are positive, be equal across protected groups (Corbett-Davies & Goel, 2018). For example, in order to satisfy false positive classification parity, the Allegheny County child neglect prediction algorithm must make an incorrect positive prediction (i.e., predict the presence of child abuse in a family where none is occurring) at the same rate for both White and Black families. Another commonly referenced notion of fairness is ”calibration,” which requires that outcomes be independent of protected class status after controlling for estimated risk (Corbett-Davies & Goel, 2018). If the aforementioned algorithm were to satisfy calibration, child abuse must be found to actually occur at similar rates in White and Black families predicted to have a 10% risk of child neglect.

These definitions may sound like they measure roughly similar phenomena, but recent research on algorithmic fairness shows that they are often in competition, producing provable mathematical tradeoffs among each other (Corbett-Davies & Goel, 2018). Optimizing calibration, for example, may result in reductions in classification parity. ProPublica's analysis of the use of COMPAS at the pretrial stage in Broward County, Florida revealed that the algorithm yielded much higher false positive rates for Black defendants than it did for White ones (Angwin, Larson, Mattu, & Kirchner, 2016), but at the same time, individuals given the same COMPAS risk score recidivated at the same rate (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017). In other words, the algorithm was calibrated, but was more likely to incorrectly classify Black defendants as ”high risk” for recidivism than White defendants. To further complicate the notion of discrimination, the algorithm used in Allegheny County to predict risk of child neglect was miscalibrated in a way that disfavored White children: White children who received the same risk score for neglect as Black children were actually less likely to be experiencing maltreatment (Chouldechova, Benavides-Prado, Fialko, & Vaithianathan, 2018). In this case, Eubanks' critiques of the algorithm's inputs and other researchers' empirically measured calibration result in directly opposing views of which racial group is experiencing discrimination.

Without a single, legally-codified definition of fairness, we see the first obstacle to a successful disparate impact claim: a recipient can argue that no less discriminatory alternative exists, since any alternative will likely involve tradeoffs across different measures of fairness. Moreover, we suggest that it is insufficient to choose one measure of fairness as the priority in all cases, since the societal costs associated with different fairness measures varies across specific applications (Corbett-Davies & Goel, 2018). For example, one might argue that the societal and/or moral cost of incorrectly detaining a Black individual who will not recidivate is far greater than the cost of incorrectly flagging a Black household for child abuse. Another person might take the opposite position, but in either case, blindly prioritizing false positive parity across both tasks would fail to recognize the unique costs associated with each one.

There also exist practical legal challenges and ambiguity regarding the existence of a less discriminatory alternative. In the realm of Title VII, scholars disagree about whether ”refusal” to adopt a less discriminatory procedure means that the employer cannot be held liable until it has actively investigated such an alternative and subsequently rejected it (Barocas & Selbst, 2016). This debate raises the question of whether employers should be held responsible to perform a costly, exhaustive search of all potential alternatives, or whether the cost of doing such a search would functionally mean that less discriminatory alternatives do not exist. According to the U.S. Department of Justice's guidance regarding Title VI, the burden is on the complainant to identify less discriminatory alternatives (U.S. Department of Justice, 2019). This may pose a significant challenge to complainants, as they may not have access to the documents and data needed to show which alternatives would be equally effective in practice.

Challenge 2: Substantial legitimate justification

The second failure mode for a disparate impact claim is that the recipient has articulated a ”substantial legitimate justification” for the challenged policy (42 U. S. C. § 2000d). As the Justice Department discloses in its Title VI legal manual, ”the precise nature of the justification inquiry in Title VI cases is somewhat less clear in application” (U.S. Department of Justice, 2019). For example, the EPA stated in its 2000 Draft Guidance for Investigating Title VI Administrative Complaints that the ”provision of public health or environmental benefits...to the affected population” was an ”acceptable justification” (Draft Title VI Guidance, 2000). This document was compiled after a 60-day period of 7 public listening sessions at the request of state and local officials seeking clarification in an effort to avoid Title VI violations (Mank, 2000). In contrast, Title VII substitutes the ”legitimate justification” requirement with a ”business necessity” stipulation (42 U. S. C. § 2000d). Because Title VI covers a broad scope of federally funded programs, ”legitimate justification” must be defined on a case-by-case basis, whereas ”business necessity” has a narrower meaning in case law due to Title VII’s specific focus on hiring practices (U.S. Department of Justice, 2019).

In the case of programmatic decision-making, discrimination may occur when practitioners do not properly audit their algorithm before and while it is deployed. Such an audit could take many forms, such as running a randomized control trial before permanently implementing an algorithm or releasing public reports every year regarding how well the algorithm is performing. (For the purposes of the following discussion, we assume that the task at hand is one of binary/multiclass classification, also known as a ”screening procedure”). In the field of machine learning, algorithms are commonly trained by iteratively improving performance on a given dataset, as measured by average classification accuracy (Alpaydin, 2009). If average classification accuracy is not disaggregated across protected groups present in the dataset, disparities in the algorithm’s performance may only be discovered once the algorithm is already deployed for real-world use (Buolamwini & Gebru, 2018), which could result in a subsequent disparate impact claim. In this sequence of events, the potentially offending entity was optimizing for overall accuracy and failed to take the possibility of disparate impact into account.

This scenario raises the question of whether the desire to optimize raw predictive accuracy counts as a ”substantial legitimate justification” for an algorithm whose outputs are biased. It seems plausible that any recipient could argue that predictive accuracy is a legitimate justification: after all, optimizing accuracy maximizes the total number of decisions made correctly, given that the demographic makeup of the dataset resembles that of the real-world population. Optimizing for any other metric, such as an arbitrary fairness measure, may lead to an algorithm with lower overall predictive accuracy (Zliobaite, 2015; Kleinberg, Mullainathan, & Raghavan, 2016). A recipient of a disparate impact claim could argue that maximizing accuracy leads to higher efficiency and lower costs for cash-strapped government agencies. In the Allegheny County example, having an algorithm accurately flag families for risk of child neglect reduced the time required to manually screen applications, saving time and labor. Because ”substantial legitimate justification” is relatively ambiguous and case-specific, it may be difficult for a complainant to prove that maximizing classification accuracy is not a legitimate justification.

Challenge 3: A disparate impact and disparate treatment Catch-22

It’s important to note that optimizing accuracy and fairness measures is not always a zero-sum game. In the aforementioned research about gender in criminal risk assessment, including gender as a variable in the dataset improved calibration and predictive accuracy because women with similar criminal histories to men recidivate at lower rates (Skeem et al., 2016) (Notably, gender is not a protected attribute under disparate impact clauses in civil rights law). Similarly, in other cases, we may be able to improve predictive accuracy and produce gains in fairness measure(s) if some predictive latent variable is identified and included in the dataset (Jung, Corbett-Davies, Shroff, & Goel, 2018).

Consider the case of a hypothetical algorithm that estimates recidivism risk and takes race as an input, but does not take criminal history as an input. Assume in this scenario that criminal history is more predictive of recidivism than race. If Black people are disproportionately likely to have prior convictions - perhaps due to disparate policing practices - then the algorithm will ”penalize” all Black people by giving them higher risk scores, even ones without prior convictions. If criminal history is added to the dataset and the algorithm is retrained, the algorithm's accuracy will increase due to the addition of a predictive variable. In addition, the algorithm's performance on fairness measures may increase as well, since Black people without criminal histories will no longer receive a penalty for their racial status.

It may be the case, however, that the latent variable whose inclusion would improve fairness and accuracy is the protected attribute itself (Jung et al., 2018). Including gender as an input to the algorithm would resolve the unequal outcomes in which women are unfairly penalized, but at the same time, explicitly altering decisions based off of an individual's gender is a clear example of disparate treatment (42 U. S. C. § 2000d). The same would be true with regard to protected attributes under Title VI such as race, national origin, and religion. Disparate treatment, in which policies explicitly treat members of different protected groups differently, is prohibited by Title VI, as well as many other civil rights laws (U.S. Department of Justice, 2019). Disparate treatment cases are arguably easier to prove, since discrimination is explicitly codified in a recipient's policies, while disparate impact cases rely on measures of a policy's outcomes de facto (Selmi, 2005). The fact that both disparate treatment and disparate impact violate civil rights statutes may create a Catch-22 for entities seeking to resolve disparate impact in algorithmic decision-making.

Indeed, Kroll et al. (2016) note this tension as manifested in the Supreme Court's decision in a 2009 case involving Title VII, *Ricci v. DeStefano* (2009). In the case, the New Haven Civil Service Board (CSB) refused to certify the results of a facially neutral test for firefighter promotions out of disparate impact concerns, noting that the pass rate for minorities was half that for whites. As Kroll et al. (2016) note, the Court's decision to rule *against* the CSB ”demonstrates the tension between disparate treatment and disparate impact,” since a neutral policy can create disparate outcomes, but mitigating the disparate impact would require discriminatory treatment of different groups.

Remedies

As we have seen from the above analysis, there is reason to believe that today's concerns regarding algorithmic bias will not be resolved in the courts alone, despite the high number of pending court cases regarding the use of algorithms. In the Constitutional realm, absent a suspect classification, both disparate impact and discriminatory intent are needed to prove a violation of the law. In addition, the current requirements to make a successful claim of disparate impact under civil rights law are vague with regards to defining what a discriminatory outcome is, which may allow recipients of complaints to leverage whichever mathematical constructs of fairness best support the use of their algorithm.

If we cannot expect to find remedies from the judiciary, where should citizens turn for relief? To address the above concerns, we propose a remedy in the form of a unified, collaborative effort between the agencies and legislatures, both at the federal and state levels. We detail what such an effort would look like below, using an international regulation to inform our proposals.

The European Union's General Data Protection Regulation (GDPR) offers a compelling case for broad legal regulations coupled with significant enforcement power. The GDPR provides strong protections for individual privacy by allowing governmental agencies to pursue fines and investigations into private companies for data mismanagement and privacy breaches (Steinhardt, 2018). With regards to automated decision-making, the GDPR (2016) makes mention of a ”right to explanation” for users who seek explanation for decisions made about them (e.g., loan denials) (Goodman & Flaxman, 2017). One of the European Commission's senior advisory bodies on data protection released a set of guidelines regarding automated decision-making, which included requirements for companies to provide explanations for how users' personal data was used by the algorithm (Casey, Farhangi, & Vogl, 2018). The same body even included a recommendation for companies to introduce ”procedures and measures to prevent...discrimination” and to perform ”frequent assessments...to check for any bias” (17/EN. WP 251, 2017).

The fact that the mandates behind the GDPR have been enforced in practice leads us to suggest an approach in the U.S. that similarly combines comprehensive legislation with new enforcement powers for government agencies (Lawson, 2019). Of course, attitudes and policies regarding the regulation of private companies differ in the U.S. and the EU (Hawkins, 2019). Thus, our proposal would not seek to impose regulations on all private companies across the US, but rather public entities that are already subject to significant government oversight, such as federal agencies or federally-funded programs. Indirectly, this implicates private companies such agencies may contract with to provide tools or services in their use of algorithmic technology.

The remedies we suggest apply to both of the main use cases we previously described; for federally funded agencies, these remedies may be enacted through legislation or executive rule-making. Similarly, these remedies could also be applied at the state and local level. In both cases, we recommend the creation or significant expansion of agencies focused specifically on the technical oversight and evaluation of algorithmic tools. For example, such an existing agency that might take up this burden could be the newly created Science, Technology Assessment, and Analytics team at the U.S. Government Accountability Office (U.S. Government Accountability Office, 2019). While courts have been reluctant to conduct a ”searching analysis of alternatives,” federal agencies are ”subject matter experts charged with Title VI enforcement duties” and ”are well-equipped to...evaluate carefully potential less discriminatory alternatives” (U.S. Department of Justice, 2019).

Our remedy additionally attempts to recognize and address the significant gap in current civil rights legislation with regards to definitions of discriminatory intent and disparate impact — which can generate a Catch-22 of sorts, even for well-meaning actors. Existing civil rights legislation largely focuses on barring discriminatory intent that results in differential treatment on the basis of protected attributes, such as race. Today, however, we see that in order to remedy unintended discrimination in algorithmic decision-making, we may have to take into account such protected attributes: essentially, using differential treatment to ameliorate disparate outcomes. Federal and state legislation must acknowledge this nuance, allowing practitioners to use protected attributes data to promote the most fair outcomes, where the relevance of such data and a suitable notion of fairness are determined on a case by case basis. For example, under a bail reform law in New Jersey, agencies may collect information about a defendant’s race and gender for potential use in a risk assessment calculation, subject to the condition that decisions are not discriminatory along race or gender lines (NJ Rev Stat § 2A:162-25, 2014).

Legislation (or other regulation) should stipulate that public agencies that are going to adopt algorithms to help make decisions must submit the following information to a relevant oversight agency (at the federal level, the office described earlier, and at the state level, some state or local agency with relevant expertise) prior to the algorithm’s adoption:

- *What decision will the algorithm be used to make or help make?* How was that decision or type of decision made before the use of the algorithm?
- *What are the reasons to implement such an algorithm?* Is the algorithm less expensive, or will it increase efficiency? Is the intent to make the decision-making process more objective?
- *What are the particular use cases and use context of the algorithm?* How will the algorithm’s outputs be interpreted? Will a human decision-maker be involved? Who is the population that the algorithm may be used on? Are there any exceptions to this policy?
- *How will the algorithm be evaluated and, if necessary, revised?* Has funding been allocated for regular oversight? Who will be performing the evaluations and how? Is the text (or source code) or training data of the algorithm publicly available?
- *Were alternatives considered?* What other options were considered, and why was this

one chosen? What were tradeoffs between the different choices?

The submission of this information to a governmental body and the public before an algorithm is employed in practice could provide greater clarity to both the public and regulators regarding discriminatory intent and the potential for discriminatory outcomes. Furthermore, by actively requiring actors to come up with a plan to monitor the algorithm, consider alternatives, and think critically about the algorithm in the context of human systems, this policy may decrease the likelihood of algorithms producing unintended negative consequences in practice.

Acknowledgments

We would like to thank Keith Schwarz for helpful feedback.

References

17/EN. WP 251. (2017). *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*

ACLU, Outten & Golden LLP, and the Communications Workers of America. (2019). *Facebook EEOC complaints.* https://www.aclu.org/cases/facebook-eeoc-complaints. (Online; accessed June 1, 2019)

Alexander v. Choate, 469 U.S. 287 (1985)

Allegheny County. (2019). *DHS funding.* https://www.county.allegheny.pa.us/Human-Services/About/Funding-Sources.aspx. (Online; accessed May 18, 2019)

Alpaydin, E. (2009). *Introduction to machine learning.* MIT press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, May*, *23*.

Arlington Heights v. Metropolitan Housing Dev. Corp., 429 U.S. 252 (1977)

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).

Casey, B., Farhangi, A., & Vogl, R. (2018). Rethinking explainable machines: The GDPR's "right to explanation" debate and the rise of algorithmic audits in enterprise.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency* (pp. 134–148).

Civil Rights Act of 1964 Title VI, 78 Stat. 252, 42 U. S. C. § 2000d.

Civil Rights Act of 1964 Title VII, 42 U.S.C. §2000e et seq.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797–806).

Craig v. Boren, 429 U.S. 190 (1976))

Desmarais, S. L., & Lowder, E. M. (2019). Pretrial risk assessment tools: A primer for judges, prosecutors, and defense attorneys. *MacArthur Foundation Safety and Justice Challenge*.

Draft Title VI Guidance for EPA Assistance Recipients Administering Environmental Permitting Programs (Draft Recipient Guidance) and Draft Revised Guidance for Investigating Title VI Administrative Complaints Challenging Permits (Draft Revised Investigation Guidance); Notice, 65 Fed. Reg. 124 (June 27, 2000). Federal Register: The Daily Journal of the United States.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

Fishel, S., Flack, D., & DeMatteo, D. (2018). Computer risk algorithms and judicial decision-making. *Monitor on Psychology*.

Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*.

Giammarise, K. (2017). *Allegheny County DHS using algorithm to assist in child welfare screening.* https://www.post-gazette.com/local/region/2017/04/09/Allegheny-County-using-algorithm-to-assist-in-child-welfare-screening/stories/201701290002. (Online; accessed May 18, 2019)

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine*, *38*(3), 50–57.

Griggs v. Duke Power Co., 401 U.S. 424 (1971)

Hawkins, D. (2019). *The cybersecurity 202: Why a privacy law like gdpr would be a tough sell in the U.S.* https://www.washingtonpost.com/news/powerpost/paloma/the-cybersecurity-202/2018/05/25/the-cybersecurity-202-why-a-privacy-law-like-gdpr-would-be-a-tough-sell-in-the-u-s/5b07038b1b326b492dd07e83/?utm_term=.1cc41e57f9cf. (Online; accessed May 18, 2019)

Hurley, D. (2018). Can an algorithm tell when kids are in danger. *New York Times*, *2*.

Jung, J., Corbett-Davies, S., Shroff, R., & Goel, S. (2018). Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. *U. Pa. L. Rev.*, *165*, 633.

K.W. v. Armstrong, No. 14-35296 (9th Cir. 2015)

Latessa, E., Smith, P., Lemke, R., Makarios, M., & Lowenkamp, C. (2009). Creation and validation of the Ohio risk assessment system: Final report. *Cincinati, OH: University of Cincinnati*.

Lau v. Nichols, 414 U.S. 563 (1974)

Lawson, R. P. (2019). *GDPR enforcement actions, fines pile up.* https://www.manatt.com/Insights/Newsletters/Advertising-Law/GDPR-Enforcement-Actions-Fines-Pile-Up. (Online; accessed May 18, 2019)

Lecher, C. (2018). What happens when an algorithm cuts your health care. *The Verge*.

Legal Information Institute. (2018a). *Equal protection.* https://www.law.cornell.edu/wex/equal_protection. (Online; accessed May 18, 2019)

Legal Information Institute. (2018b). *Intermediate scrutiny.* https://www.law.cornell.edu/wex/intermediate_scrutiny. (Online; accessed June 1, 2019)

Lewis, N. (2018). *Will AI remove hiring bias?* https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/will-ai-remove-hiring-bias-hr-technology.aspx. (Online; accessed May 18, 2019)

Mank, B. C. (2000). The draft recipient guidance and the draft revised investigation guidance: Too much discretion for epa and a more difficult standard for complainants? *Environmental Law Reporter*, *30*.

Misra, T. (2018). *When criminalizing the poor goes high-tech.* https://www.citylab.com/equity/2018/02/the-rise-of-digital-poorhouses/552161/?platform=hootsuite. (Online; accessed May 18, 2019)

NJ Rev Stat § 2A:162-25 (2014)

O.J. (L 119) (2016). *Reg (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Dir 95/46/EC (General Data Protection Regulation)*

Personnel Adm’r of Massachusetts v. Feeney, 442 U.S. 256 (1979)

Reg (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free move-

ment of such data, and repealing Dir 95/46/EC (General Data Protection Regulation). (2016).

Ricci v. DeStefano, 557 U.S. 557 (2009)

Selmi, M. (2005). Was the disparate impact theory a mistake. *Ucla L. Rev.*, *53*, 701.

Skeem, J., Monahan, J., & Lowenkamp, C. (2016). Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior*, *40*(5), 580.

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stan. L. Rev.*, *66*, 803.

State v. Loomis, 881 N.W.2d 749 (2016)

Steinhardt, E. (2018). *European regulators are intensifying GDPR enforcement.* https://www.insideprivacy.com/eu-data-protection/european-regulators-are-intensifying-gdpr-enforcement/. (Online; accessed May 18, 2019)

Summers, C., & Willis, T. (2010). Pretrial risk assessment research summary. *Washington, DC: Bureau of Justice Assistance*.

United States. Department of Justice. (2019). *Title VI Legal Manual (Updated)*

U.S. Government Accountability Office. (2019). *Our new science, technology assessment, and analytics team.* https://blog.gao.gov/2019/01/29/our-new-science-technology-assessment-and-analytics-team/. (Online; accessed May 18, 2019)

U.S. Const. amend. V.

U.S. Const. amend. XIV.

VanNostrand, M., et al. (2009). Pretrial risk assessment in Virginia.

Weinzweig, M. J. (1983). Discriminatory impact and intent under the equal protection clause: The Supreme Court and the mind-body problem. *Law & Ineq.*, *1*, 277.

Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*.

Marissa Gerchick is a rising senior at Stanford University studying Mathematical and Computational Science. She is interested in using data-driven tools to improve the American criminal justice system.

Matthew Sun is a rising senior at Stanford double majoring in Computer Science and Public Policy. He leads a student group called CS+Social Good and is interested in applied AI research for socially relevant issues.

What Metrics Should We Use To Measure Commercial AI?

Cameron Hughes (Northeast Ohio ACM Chair; cameronhughes@acm.org)
Tracey Hughes (Northeast Ohio ACM Secretary; tracey.hughes@neoacmchapter.org)

DOI: 10.1145/3340470.3340479

In AI Matters Volume 4, Issue 2, and Issue 4, we raised the notion of the possibility of an AI Cosmology in part in response to the "AI Hype Cycle" that we are currently experiencing. We posited that our current machine learning and big data era represents but one peak among several previous peaks in AI research in which each peak had accompanying "Hype Cycles". We associated each peak with an epoch in a possible AI Cosmology. We briefly explored the logic machines, cybernetics, and expert system epochs. One of the objectives of identifying these epochs was to help establish that we have been here before. In particular we've been in the territory where some application of AI research finds substantial commercial success which is then closely followed by AI fever and hype. The public's expectations are heightened only to end in disillusionment when the applications fall short. Whereas it is sometimes somewhat of a challenge even for AI researchers, educators, and practitioners to know where the reality ends and hype begins, the layperson is often in an impossible position and at the mercy of pop culture, marketing and advertising campaigns. We suggested that an AI Cosmology might help us identify a single standard model for AI that could be the foundation for a common shared understanding of what AI is and what it is not. A tool to help the layperson understand where AI has been, where it's going, and where it can't go. Something that could provide a basic road map to help the general public navigate the pitfalls of AI Hype.

Here, we want to follow that suggestion with a few questions. Once we define and agree on what is meant by the moniker artificial intelligence and we are able to classify some application as actually having artificial intelligence, another set of questions immediately present themselves:

- How intelligent is any given AI application?
- How much intelligence does any given AI application have?
- How much intelligence does an application need to be classified as an AI application?
- How reliable is the process that produced the intelligence for any given AI application?
- How transparent is the intelligence in any given AI application?

The answers to these questions require some kind of qualitative and quantitative metrics. Namely, how much intelligence does any given AI application have and what is the quality of that intelligence. Further, how could we congeal the answers to these questions so that they can be used to capture (in label form) the 'AI Ingredients' of any technology aimed at the general public?

The amount of education is often used as one metric for intelligence. We refer to individuals as having a grade school, high school or college-level education. Could we employ a similar metric for AI applications? Would it be feasible to classify AI applications in terms of grade levels? For instance, an AI application with grade level 5 would be considered to have more intelligence than a 4th grade AI application. Any application that didn't meet 1st grade level would not be considered an AI application and applications that achieved better than 12th grade would be considered advanced AI applications. But how could we determine the grade level of any given AI application?

A consensus set of metrics that could be passed on to the general public has not yet prevailed. In this AI hype cycle, if an application uses any artifact from any AI technique a vendor is quick to advertise it as an AI application. It would be useful to have a metric that would indicate exactly how much AI is in the purported application. We have this kind of information for other products e.g. how much chocolate is actually in the chocolate bar or how much real fruit is actually in the fruit juice being sold to the consumer. Exactly how much AI does that drone have? Or how much AI is actually in that new social media application? The 'how much' question requires a quantitative metric of some sort and the grade of intelligence involved requires the qualitative metric. What if applications that claimed to be AI capable were required to state the metrics on the label or in the advertising? For example, A vendor might state: “Our new social media application is 2% level 3 AI!” This kind of simple metric scheme would help to mitigate AI Hype cycles.

In addition to characterizing the quantity and quality of the embedded AI, requiring a reliability metric like MTBAIF (Mean Time Between AI Failure) is also desirable. Stating how much intelligence is in an application and what grade level of intelligence is in an application provides a good start. However, the reliability of the AI (i.e its limits, tolerances, certainty, etc.) and a transparency metric that indicates the ontology, inference predisposition/bias, and type and quality of knowledge would give the user some real indication of the utility of the application.

Knowledge Ingredients

What if we could provide 'Knowledge Ingredient' Labels for our AI-based hardware/software technologies like those shown in Figure 1.

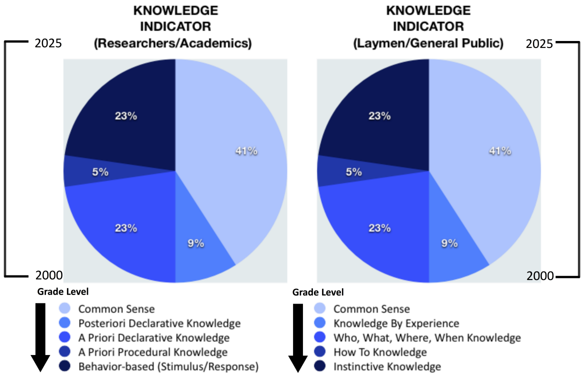


Figure 1: Pie charts for a Knowledge Indicator that reflects the Knowledge Ingredients of an AI system.

In Figure 1, the percentage of the types of knowledge the system is comprised of is represented in a pie chart. The color indicates the grade levels from lower to higher. In this case, the system has 41% of the knowledge is ”Common Sense” with a relative low grade level as compared to the high grade level of ”Instinctive Knowledge” at 23%. The expiration date indicates the time frame for the viability of the knowledge from 2000 - 2025. Outside of the indicated time frame, the knowledge would be consider obsolete. There are two indicators, one for researchers and practitioners and the other for the laymen. The difference in the indicators is the terminology used to describe the types of knowledge. Here is our lists of metrics:

1. Percentage of software/hardware dedicated to the implementation of AI techniques
2. Grade Level
3. Reliability (limits, tolerances, certainty, MTBAIF)
4. Transparency (predisposition/bias)

AI Metrics

The notion of metrics to measure AI performance has been under active investigation since the very beginnings of AI research and a standard widely used and accepted set of metrics remain elusive. The PerMIS (Performance Metrics for Intelligent System) workshops were started in 2000. The PerMIS workshops are dedicated to:

> *“...defining measures and methodologies of evaluating performance of intelligent systems started in 2000, the PerMIS series focuses on applications of performance measures to applied problems in commercial, industrial, homeland security, and military applications.”*

The PerMIS workshops were originally cosponsored by SIGART (now SIGAI), NIST, IEEE, ACM, DARPA, and NSF. These workshops endeavored to identify performance intelligence metrics in many areas such as: ontologies, mobile robots, intelligence interfaces, agents, intelligent test beds, intelligent performance assessment, planning models, autonomous systems, learning approaches, and embedded intelligent components. ALFUS (Autonomous Levels For Unmanned Systems) defines a framework for characterizing human interaction, the mission and environmental complexity of a ”system of systems”. The purpose of ALFUS was to determine the level of autonomy, a necessary but not sufficient component of an intelligent system. According to Ramsbotham [1]:

> *“Intelligence implies an ability to perceive and adapt to external environments in real-time, to acquire and store knowledge regarding problem solutions, and to incorporate that knowledge into system memory for future use.”*

describes the autonomous intelligent systems of systems based on the 4D/RCS Reference Model Architecture for Learning developed by Intelligent Systems Division of the NIST since 1980s. In order to attempt to develop some type of metric, the functions that comprised the intelligent behavior had to be decomposed into specific hardware and software characteristics. This was called SPACE (Sense, Perceive, Attend, Apprehend, Comprehend, Effect action):

- **Sense:**
  To generate a measurable signal (usually electrical) from external stimuli. A sensor will often employ techniques (for examples, bandpass filtering or thresholding) such that only part of the theoretical response of the transducer is perceived.
- **Perceive:**
  To capture the raw sensor data in a form (analog or digital) that allows further processing to extract information. In this narrow construct perception is characterized by a 1:1 correspondence between the sensor signal and the output data.
- **Attend:**
  To select data from what is perceived by the sensor. To a crude approximation, analogous to feature extraction.
- **Apprehend:**
  To characterize the information content of the extracted features. Analogous to pattern recognition.
- **Comprehend:**
  To understand the significance of the information apprehended in the context of existing knowledge–in the case of automata, typically other information stored in electronic memory.
- **Effect action:**
  To interact with the external environment or modify the internal state (e.g., the stored information comprising the ”knowledge base” of the system) based on what is comprehended.

The purpose or “collective mission performance” of a systems of systems was also categorized based on functional and architectural complexity. The mission and environmental complexity and the degree of human interaction determining the level of autonomy could be applied to these categories [1]:

1. **Leader-Follower**
   Intelligent behavior exhibited by single node, and replicated (sometimes with minor adaptation) by other nodes.
2. **Swarming (simple)**
   Loosely structured collection of interacting agents, capable of moving collectively.
3. **Swarming (complex)**
   Loosely structured collection of interacting agents, capable of individuated behavior to effect common goals.
4. **Homogenous intelligent systems**
   A relatively structured collection of identical (or at least similar) agents, wherein collective system performance is optimized by optimizing the performance of individual agents.
5. **Heterogeneous intelligent systems**
   A relatively structured heterogeneous collection of specialized agents, wherein the functions of intelligence distributed among the diverse agents to optimize performance of a defined task or set of tasks.
6. **Ad hoc intelligent adaptive systems**
   A relatively unstructured and undefined heterogeneous collection of agent, wherein the functions comprising intelligence are dynamically distributed across the system to adapt to changing tasks.

From less complex groups with low mission and environmental complexity and high human interaction (like Leader-Follower) to the most sophisticated high mission and environmental complexity and no human interaction (like Ad Hoc Intelligent Adaptive Systems), these categories of systems are in a somewhat order. Figure 2 shows where the extremes of these categories could be graphed.

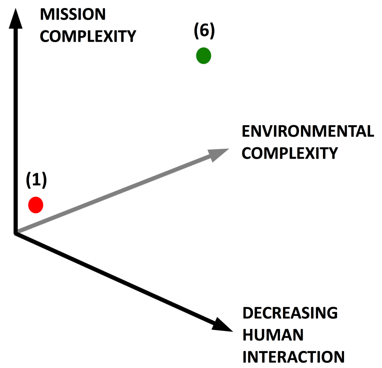


Figure 2: Lead Follower and Ad Hoc Intelligent Adaptive Systems are located on the 3D space of Complexity and Human Interaction.

Ultimately, Ramsobotham [1] concluded:

> *"Even given a comprehensive framework, as we begin to build more complex intelligent systems of systems, we will need to acquire knowledge and improve analytic tools and metrics. Among the more important will be: ... Better models and metrics for characterizing limits of information assurance based on these effects. This will be both a critical need and a major challenge."*

The most known metric used for researchers and commercial AI applications has been the Turing Test developed by Alan Turing in 1950. The purpose of the test was to evaluate a machine's ability to demonstrate intelligent behavior. That behavior was to be indistinguishable from a human being exhibiting the same behavior. A human judge was to evaluate a natural language conversation between a human and the potential "intelligent machine" shown in Figure 3.

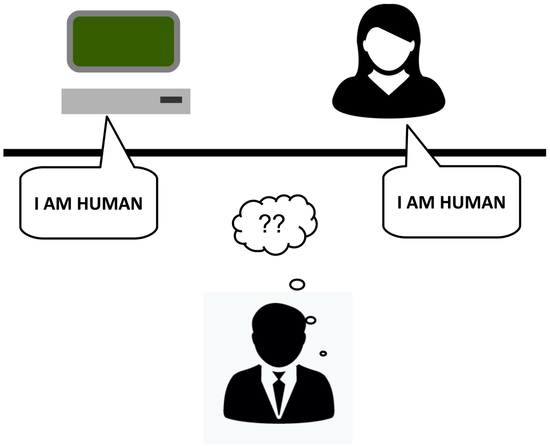


Figure 3: Human Evaluator was to determine which was not human.

Is this test actually a metric for "intelligence"? This has been debated for many years. If a NLG agent can produce responses that simulate human responses does that mean that the system is intelligent? Probably not, but that has not stopped the use of the Turing Test or AI software developers heralding that their systems has passed or almost passed the Turing Test. Now passing the test includes simulating the human voice like the Google Duplex AI Assistant.

In "Moving Beyond the Turing Test with the Allen AI Science Challenge" [2], the authors describe the Allen AI Science Challenge "Turing Olympics", a series of tests that explore many capabilities that are considered associated with intelligence. These capabilities include language understanding, reasoning, and commonsense knowledge needed to perform smart or intelligent activities. This would replace the Alan Turing pass/fail model. The idea of the Challenge was to have a four-month-long competition where researchers were to build an AI agent that could answer an eighth-grade multiple choice science questions. The AI would demonstrate its ability to utilize state-of-the-art natural language understanding and knowledge-based reasoning. Below summarizes the nature of the competition:

Number of total (4) choice questions: **5,083**

Training set Questions: **2,500**

Validation set confirming model performance: **8,132**

Legitimate questions: **800**

Final Test + validation set for final score: **21,298**

Final Legitimate Questions: **2,583**

Baseline Score random guessing: **25%**

The final results of the test, the top three competitors had scores with a spread of 1.05% with the highest score 59.31% which is considered a failing grade. Each of the winning model utilized standard information-retrieval-based methods which were not able to pass the eighth grade science exams. What is required is:

> *" ...to go beyond surface text to a deeper understanding of the meaning underlying each question, then use reasoning to find the appropriate answer."*

In this case, such a system would have a 1st grade level Knowledge Quality based on our quasi Knowledge Ingredient Indicator. Based on the article [2]:

> *"All three winners said it was clear that applying a deeper, semantic level of reasoning with scientific knowledge to the questions and answers would be the key to achieving scores of 80% and higher and demonstrating what might be considered true artificial intelligence."*

Here we've provided a very cursory look at a very limited set of possibilities for measuring commercial AI. In the next issue, will go a little further and dig a little deeper into the question of how do we communicate basic AI metrics and ingredients for commercial AI to the layperson.

References

[1] Ramsbotham, Alan.J. (2009). *Collective Intelligence: Toward Classifying Systems of Systems*. PerMIS'09.

[2] Schoenick, Carissa, Clark, Peter, T. et.al (2017). *Moving Beyond the Turing Test with the Allen AI Science Challenge*. CACM, VOL.60 (N0.9), pp. 60-64.

Cameron Hughes is a computer and robot programmer. He is a Software Epistemologist at Ctest Laboratories where he is currently working on A.I.M. (Alternative Intelligence for Machines) and A.I.R (Alternative Intelligence for Robots) technologies. Cameron is the lead AI Engineer for the Knowledge Group at Advanced Software Construction Inc. He is a member of the advisory board for the NREF (National Robotics Education Foundation) and the Oak Hill Robotics Makerspace. He is the project leader of the technical team for the NEOACM CSI/CLUE Robotics Challenge and regularly organizes and directs robot programming workshops for varying robot platforms. Cameron Hughes is the co-author of many books and blogs on software development and Artificial Intelligence.

Tracey Hughes is a software and epistemic visualization engineer at Ctest Laboratories. She is the lead designer for the MIND, TAMI, and NOFAQS projects that utilize epistemic visualization. Tracey is also a member of the advisory board for the NREF (National Robotics Education Foundation) and the Oak Hill Robotics Makerspace. She is the lead researcher of the technical team for the NEOACM CSI/CLUE Robotics Challenge. Tracey Hughes is the co-author with Cameron Hughes of many books on software development and Artificial Intelligence.

Crosswords

Adi Botea (IBM Research, Ireland; adibotea@ie.ibm.com)

DOI: 10.1145/3340470.3340480

1		2	3	4	5	■	6	7	8	9	10	11
	■	12				■	13					
14	15	A				■	16					
17			I	■	18	19			■	20		
21			■	M				■	23			
24			25		A		■	26				
■	■	■	27			T	28			■	■	■
29	30	31			■	32	T			33	34	35
36				■	37			E	■	38		
39			■	40				■	R			
42			43			■	44	45		S		
46						■	47				■	
48						■	49					

Across: 1) French city by the Straight of Dover **6)** Go through a printed paper **12)** A point in time **13)** Protected with a concrete defense **14)** A.C. ___, from Saved by the Bell **16)** Seas at a raised level **17)** ___ Braxton, American singer **18)** Undesired spot **20)** Gear tooth **21)** Be indebted **22)** Mine in France **23)** Ireland in the local language **24)** Buckingham guard attire (2 wds.) **26)** Summation circuit **27)** Crying out loud **29)** Follow up actively (2 wds.) **32)** Said more formally **36)** Nominated as a fellow **37)** Continuous pain **38)** Output of a mining procedure **39)** Verb invoked with ability **40)** Legal argument **41)** Old tourist attraction **42)** Unpleasant experience **44)** Clothes area **46)** ___ Wonder from the world of music **47)** Person hired to help **48)** Unexcitingly **49)** Present on the list of requirements

Down: 1) ___ and Pollux, the Gemini **2)** Given away temporarily **3)** Against **4)** Curling surface **5)** Croatian neighbor **6)** Lose consciousness **7)** Diplomatic skill **8)** Pale at a party **9)** Tranquil **10)** Poem by Edgar Allan Poe **11)** ___ Dijkstra, the famous computer scientist **15)** Prussian lion **19)** Salad ingredient **22)** React with no joy **23)** ___ computing in IoT **25)** Farm production **26)** Before... the prefix **28)** Resident near Cornell University **29)** Approach with no good promise **30)** Medicine on paper **31)** Binary string? **33)** Returned from a dream experience **34)** Great Lake **35)** Marked from an impact **37)** Park way **40)** Bucket **41)** Entered a bike race **43)** First lady **45)** Metaphoric curtain in front of reality

Previous puzzle solution: ASNEER - ALISTS - SCARCE - BERTIE - LAMEST - BAREST - ORE - RAINIEST - PEI - SANE - TRUE - ESTONIA - MASER - HANGMAN - CRAMP - RAGTOPS - HARM - BARI - CRY - INTERIMS - TIS - ROUTER - BEZANT - PUREED - ATONCE - STORKS - REDEEM

Acknowledgment: I thank Karen McKenna for her feedback. The grid is created with the AI system Combus (Botea, 2007).

References

Botea, A. (2007). Crossword Grid Composition with A Hierarchical CSP Encoding. In *Proceeding of the 6th CP Workshop on Constraint Modelling and Reformulation ModRef-07.*

Adi Botea is a research scientist at IBM Research, Ireland. His interests include AI planning, heuristic search, automated dialogue systems, pathfinding, journey planning, and multi-agent path planning. Adi has co-authored more than 70 peer-reviewed publications. He has published crosswords puzzles in Romanian, in national-level publications, for almost three decades.