acm SIGAI

AI Matters

Annotated Table of Contents

Essay Contest

Artificial Intelligence: The Societal Responsibility to Inform, Educate, and Regulate

Alexander D. Hilton

Full article: http://doi.acm.org/10.1145/3362077.3362088

Winning essay from the 2018 ACM SIGAI Student Essay Contest

The Necessary Roadblock to Artificial General Intelligence: Corrigibility

Yat Long Lo, Chung Yu Woo & Ka Lok Ng

Full article: http://doi.acm.org/10.1145/3362077.3362089

Winning essay from the 2018 ACM SIGAI Student Essay Contest

AI Fun Matters

Adi Botea

Full article: http://doi.acm.org/10.1145/3362077.3362090

AI generated Crosswords

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: 10.1145/3362077

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(3)

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: 10.1145/3362077.3362078

Issue overview

Welcome to the third issue of the fifth volume of the AI Matters Newsletter. With this issue, we want to welcome our new SIGAI Executive Committee. Elections were completed this Spring, and we have a new leadership team in place. Sanmay Das of Washington University in St. Louis (former vice-chair) is the new chair, Nicholas Mattei of Tulane University the new vice-chair, and John Dickerson of the University of Maryland the new secretary-treasurer. Nicholas and John were formerly active as the appointed AI and Society and Labor Market officers respectively. Sven Koenig will transition into the role of past-chair, and continue to serve on the EC in that role.

The new officers want to express their sincere thanks to Sven Koenig and to Rosemary Paradis (the former secretary/treasurer) for the wealth of novel initiatives they spearheaded in the last three years and the untiring energy they brought to their roles. SIGAI is deeply indebted to them!

We would like to mention that there has been a lot of activity in the space of significant awards in AI. The inaugural SIGAI Industry Award for Excellence in Artificial Intelligence (AI) was presented at IJCAI 2019. The award went to the Real World Reinforcement Learning Team from Microsoft, for identification and development of cutting-edge research on contextual-bandit learning that led to new decision support tools that were broadly integrated into a broad range of Microsoft products. John Langford and Tyler Clintworth received the award on behalf of the Microsoft team and presented a talk on the work at IJCAI. For more on this award, please see https://sigai.acm.org/awards/industry_award.html

We also congratulate Marijn Heule, Matti Järvisalo, Florian Lonsing, Martina Seidl and Armin Biere who have been awarded the 2019 IJCAI-JAIR prize for their 2015 paper "Clause Elimination for SAT and QSAT" (https://jair.org/index.php/jair/article/view/10942). This paper describes fundamental and practical results on a range of clause elimination procedures as preprocessing and simplification techniques for SAT and QBF solvers. Since its publication, the techniques described therein have been demonstrated to have profound impact on the efficiency of state-of-the-art SAT and QBF solvers. The work is elegant and extends beautifully some well-established theoretical concepts. In addition, the paper gives new emphasis and impulse to pre- and inprocessing techniques - an emphasis that resonates beyond the two key problems, SAT and QBF, covered by the authors.

We would also like to note that SIGAI and AAAI will be jointly presenting a new annual award for the best doctoral dissertation in AI. The award will be presented at AAAI, and nominations for the inaugural award are due by November 15, 2019. Please see http://sigai.acm.org/awards/nominations.html for details and information on how to submit a nomination!

This issue is full of great new articles and stories for you! We open with the annual report of SIGAI. We then bring you a story from a new way to teach kids and families about AI: Technovation Familes' AI challenge, which brings AI into the home by educating parents and children about AI and providing an opportunity for them to prototype AI solutions to real-world problems. They are seeking new mentors for this year's challenges!

In our regular articles, Michael Rovatsos reports on upcoming AI events and we have two submissions for AI Education. First, Michael Guerzhoy talks about building a fake news detector. Second, Marion Neumann talks about bringing AI and ML to a younger audience, much like CS for all, instead of focusing on seniors and graduate students. In the policy column, Larry Medsker summarizes recent policies covering face recognition (how much data should we record and share?), upcoming AI regulation, and more. Our final regular column is our AI crosswords from Adi Botea. Enjoy!

We have a new regular column where we invite researchers to present latest research trends in AI. In the inaugural article of this column, Tianbao Yang describes challenges and opportunities of non-convex and constrained learning.

In our contributed articles, Shari Trein et al. describe some of the opportunities and risks across four emerging AI application areas: employment, education, public safety, and healthcare, identified in a workshop with participants experiencing a range of disabilities. Finally, this issue features the second set of winning essays from the 2018 ACM SIGAI Student Essay Contest. 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.

Special Issue: AI For Social Good

Recognizing the potential of AI in solving some of the most pressing challenges facing our society, we are excited to announce that the next Newsletter of AI Matters will be a special issue on the theme of "AI for Social Good." We solicit articles that discuss how AI applications and/or innovations have resulted in a meaningful impact on a societally relevant problem, including problems in the domains of health, agriculture, environmental sustainability, ecological forecasting, urban planning, climate science, education, social welfare and justice, ethics and privacy, and assistive technology for people with disabilities. We also encourage submissions on emerging problems where AI advances have the potential to influence a transformative change, and perspective articles that highlight the challenges faced by current standards of AI to have a societal impact and opportunities for future research in this area. More details to be coming soon on http://sigai.acm.org/aimatters. Please get in touch with us if you have any questions!

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.

ACM SIGAI Activity Report

Sven Koenig (elected; ACM SIGAI Chair)
Sanmay Das (elected; ACM SIGAI Vice-Chair)
Rosemary Paradis (elected; ACM SIGAI Secretary/Treasurer)
John Dickerson (elected; ACM SIGAI Labor Market Officer)
Yolanda Gil (appointed; ACM SIGAI Past Chair)
Katherine Guo (appointed; ACM SIGAI Membership and Outreach Officer)
Benjamin Kuipers (appointed; ACM SIGAI Ethics Officer)
Iolanda Leite (appointed; ACM SIGAI Newsletter Editor-in-Chief)
Hang Ma (appointed; ACM SIGAI Information Officer)
Nicholas Mattei (appointed; ACM SIGAI AI and Society Officer)
Amy McGovern (appointed; ACM SIGAI Newsletter Editor-in-Chief)
Larry Medsker (appointed; ACM SIGAI Public Policy Officer)
Todd Neller (appointed; ACM SIGAI Education Activities Officer)
Marion Neumann (appointed; ACM SIGAI Diversity Officer)
Plamen Petrov (appointed; ACM SIGAI Industry Liaison Officer)
Michael Rovatsos (appointed; ACM SIGAI Conference Coordination Officer)
David Stork (appointed; ACM SIGAI Award Officer)

DOI: 10.1145/3362077.3362079

Abstract

We are happy to present the annual activity report of the ACM Special Interest Group on AI (ACM SIGAI), covering the period from July 2018 to June 2019.

The scope of ACM SIGAI consists of the study of intelligence and its realization in computer systems (see also its web-site at sigai.acm.org). This includes areas such as

> autonomous agents, cognitive modeling, computer vision, constraint programming, human language technologies, intelligent user interfaces, knowledge discovery, knowledge representation and reasoning, machine learning, planning and search, problem solving and robotics.

Members come from academia, industry and government agencies worldwide. ACM SIGAI recently added two new ACM SIGAI chapters, namely one professional chapter in Laguna Nigel (USA) and one student chapter at the SRM Institute of Science & Technology in Chennai.

ACM SIGAI also added two new officers this year to be able to serve its membership better, namely Iolanda Leite from the Royal Institute of Technology (Sweden) as second newsletter editor-in-chief and Marion Neumann from Washington University in St. Louis (USA) as diversity officer, thus increasing diversity in the ACM SIGAI leadership committee by increasing both the number of international officers and the number of female officers and also furthering the internationalization of the ACM SIGAI newsletter. Marion will be covering diversity also as part of the ACM SIGAI newsletter. ACM SIGAI started officers meetings at major AI conferences already in 2018 and continued the new practice in 2019 (so far at the AAAI Conference), in addition to all-officers teleconferences and a monthly ACM SIGAI leadership teleconference.

Meetings

ACM SIGAI decided to participate on a trial basis in ACM's voluntary carbon-offset program for conferences. Introduction of this scheme will give conference participants the option of making voluntary contributions to offset the carbon footprint of their trips to conferences when they register online. ACM SIGAI plans to test this scheme at upcoming editions of the AAAI/ACM AI, Ethics and Society (AIES) and ACM Intelligent User Interfaces (IUI) conferences in cooperation with AAAI and SIGCHI, respectively, and hopes that it will enable the ACM SIGAI and wider ACM membership to contribute to the environmental sustainability of our communities.

ACM SIGAI continues to support AIES, which it co-founded in 2017 to fill a scientific void. As AI is becoming more pervasive in our lives, its impact on society is more significant, raising ethical concerns and challenges regarding issues such as value alignment, safety and security, data handling and bias, regulations, accountability, transparency, privacy and workforce displacement. Only a multi-disciplinary and multi-stakeholder effort can find the best ways to address these concerns, by including experts from various disciplines, such as ethics, philosophy, economics, sociology, psychology, law, history and politics. AIES was co-located with AAAI 2019 in Honolulu and will again be co-located with AAAI 2020 in New York City.

ACM SIGAI sponsored the following conferences in addition to AIES 2019:

- WI 2018
- ASE 2018
- IVA 2018
- HRI 2019
- IUI 2019

and it will sponsor the following conferences coming up in 2019 and 2020:

- IVA 2019
- K-CAP 2019
- ASE 2019
- WI 2019
- ASE 2020
- HRI 2020
- IUI 2020

ACM SIGAI approved the following in-cooperation and sponsorship requests from events covering a wide thematic and geographical range across the international AI community:

- iWOAR 2018
- ICPRAM 2018
- IEA/AIE 2019
- FW 2018
- BIOSTEC 2019
- RecSys 2019
- FW 2019
- ICAART 2019
- AAMAS 2019
- iWOAR 2019
- KMIKS 2019
- ICPRAM 2019
- FDG 2019
- ICAIL 2019
- IC3K 2019
- IJCCI 2019
- IEA/AIE '20
- AAMAS 2020

ACM SIGAI also organizes – jointly with the Association for the Advancement of AI (AAAI) – the annual joint job fair at the AAAI conference, where attendees can find out about job and internship opportunities from representatives from industry, universities and other organizations. The AAAI/ACM SIGAI job fair was held at AAAI 2019 in Honolulu, co-organized by the ACM SIGAI labor market officer. Twenty-six employers formally attended, while a handful of exhibitors who did not formally sign up also took part. Hundreds of CVs and resumes were collected before, during and after the job fair from students, postdoctoral researchers and other job seekers via the job fair web-site; these were shared with interested employers. This year, the organizers also purchased a dedicated domain (aaaijobfair.com) to allow present and future firms and participants to view previous iterations of the job fair. The ACM SIGAI labor market officer believes that we can use insights from AI to create an even better functioning job market and works actively toward designing the job market of the future. Toward that end, he has begun to gather requirements with a large number of chairs of top computer science departments in the USA as well as in Israel and Europe and is working to formulate a model that will be translated into a larger job fair (in terms of participating employers as well as applicants) in the near future.

ACM SIGAI also co-sponsors – jointly with AAAI – the annual joint doctoral consortium at the AAAI conference, which provides an opportunity for Ph.D. students to discuss their research interests and career objectives with the other participants and a group of established AI researchers who act as their mentors. The AAAI/ACM SIGAI doctoral consortium was held at AAAI 2019 in Honolulu.

Awards

ACM SIGAI sponsors the ACM SIGAI Autonomous Agents Research Award, an annual award for excellence in research in the area of autonomous agents. The recipient is invited to give a talk at the International Conference on Autonomous Agents and Multiagent Systems (AAMAS). The 2019 ACM SIGAI Autonomous Agents Research Award was presented at AAMAS 2019 in Montreal to Carles Sierra, the vice-director of the AI Research Institute of the Spanish National Research Council, for seminal contributions to research on negotiation and argumentation, computational trust and reputation and artificial social systems.

ACM SIGAI also sponsors the ACM SIGAI Industry Award for Excellence in AI, a new annual award which is given annually to an individual or team in industry who created a fielded AI application in recent years that demonstrates the power of AI techniques via a combination of the following features: novelty of application area, novelty and technical excellence of the approach, importance of AI techniques for the approach and actual and predicted societal impact of the application. The inaugural ACM SIGAI Industry Award for Excellence in AI will be presented at the International Joint Conference on AI (IJCAI) 2019 in Macau to the Real World Reinforcement Learning Team from Microsoft for the identification and development of cutting-edge research on contextual-bandit learning, the manifest cooperation between research and development efforts, the applicability of the decision support throughout the broad range of Microsoft products and the quality of the final systems.

ACM SIGAI also recently created – jointly with AAAI – the joint AAAI/ACM SIGAI Doctoral Dissertation Award to recognize and encourage superior research and writing by doctoral candidates in AI. This new annual award will be presented at the AAAI Conference on AI in the form of a certificate and is accompanied by the option to present the dissertation at the AAAI conference as well as to submit a six-page summary to both the AAAI proceedings and the ACM SIGAI newsletter. The nomination deadline for the inaugural AAAI/ACM SIGAI Doctoral Dissertation Award will be announced later this year and is expected to be in late Fall 2019.

Public Policy Activities

ACM SIGAI promotes the discussion of policies related to AI through posts in the AI Matters blog, helps to identify external groups with common interests in AI public policy, encourages ACM SIGAI members to partner in policy initiatives with these organizations, disseminates public policy ideas to the ACM SIGAI membership through articles in the ACM SIGAI newsletter and ensures that every technologist is educated, trained and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems. ACM SIGAI participates in the ACM US Technology Policy Committee (ACM USTPC), formerly USACM, via the ACM SIGAI public policy officer and in a variety of other policy efforts, including those of other societies (such as the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems). ACM USTPC addresses US public policy issues related to computing and information technology and regularly educates and informs US Congress, the US Administration and the US courts about significant developments in the computing field and how those developments affect public policy. For example, the ACM SIGAI public policy officer joined the comments of ACM USTPC on the draft of the “20-Year Roadmap for AI Research in the US” of the Computing Community Consortium. He also studies how organizations collect and analyze data and whether these practices are consistent with recommendations by the USTPC working group on algorithmic accountability, transparency and bias. He represented ACM SIGAI via his talks “Future of Work, AI Education, and Public Policy” at EAAI 2019 and “Transparency, Accessibility, and Ethics in AI” at Dalhousie University in 2019. He was also the moderator of the panel "Are We Ready for AI" at the Annual Consumer Assembly of the Consumer Federation of America in 2019.

Educational Activities

ACM SIGAI held a second ACM SIGAI Student Essay Contest focused on AI ethics (organized by the ACM SIGAI AI and society officer), after the success of the first such competition in 2017. Students could win cash prizes of US$500 or Skype conversations with senior AI researchers from academia or industry (including the director of Microsoft Research Labs and the director of research at Google) if their essays provided good answers to one or both of the following topic areas (or any other question in this space that they considered important):

- What requirements, if any, should be imposed on AI systems and technology when interacting with humans who may or may not know that they are interacting with a machine? For example, should they be required to disclose their identities? If so, how?
- What requirements, if any, should be imposed on AI systems and technology when making decisions that directly affect humans? For example, should they be required to make transparent decisions? If so, how?

This year, ACM SIGAI received 18 submissions, of which eight were selected for publication and prizes. The winning essays are listed below in alphabetical order by author. ACM SIGAI intents to hold a third ACM SIGAI student Essay Contest later this year.

- Janelle Berscheid and Francois Roewer-Despres – *Beyond Transparency: A Proposed Framework for Accountability in Decision-Making AI Systems*
- Gage Garcia – *AI Education: A Requirement for a Strong Democracy*
- Alexander Hilton – *AI: The Societal Responsibility to Inform, Educate, and Regulate*
- Michelle Seng Ah Lee – *Context-Conscious Fairness in Using Machine Learning to Make Decisions*
- Yat Long Lo, Chung Yu Woo and Ka Lok Ng – *The Necessary Roadblock to Artificial General Intelligence: Corrigibility*
- Grace McFassel – *Embedding Ethics: Design of Fair AI Systems*
- Matthew Sun and Marissa Gerchick – *The Scales of (Algorithmic) Justice: Tradeoffs and Remedies*
- Annie Zhou – *The Intersection of Ethics and AI*

ACM SIGAI supported the "Birds of the Feather" undergraduate research challenge organized by the ACM SIGAI education officer at the Symposium on Educational Advances in AI (EAAI) 2019. Six research and liberal arts institutions participated with seven papers and one poster presentation that passed peer review. ACM SIGAI contributed US$500 of award funding for the best papers. The ACM SIGAI education officer intends to announce a Gin Rummy undergraduate research challenge at EAAI 2020.

ACM SIGAI also started discussions with the ACM Special Interest Group on Computer Science Education (ACM SIGCSE) on a collaboration for disseminating pointers to resources for AI educators and creating incentives for the production and dissemination of assignments on AI ethics.

Member Communication

ACM SIGAI communicates with its members via email announcements, the ACM SIGAI newsletter "AI Matters," the AI Matters blog and webinars:

ACM SIGAI maintains a more than 4,000 email address long mailing list for AI-related announcements to its members and friends.

ACM SIGAI publishes four issues of its newsletter AI Matters per year. The ACM SIGAI newsletter is distributed via the ACM SIGAI mailing list but also openly available on the ACM SIGAI web-site (at `sigai.acm.org/aimatters/`). AI Matters features articles of general interest to the AI community and added not only an additional editor-in-chief but also additional column editors in the past year. Recent columns, led by these and other column editors, have included AI Interviews (organized by the ACM SIGAI diversity officer), AI Amusements, AI Education (written or organized by the ACM SIGAI education officer), AI Policy Issues (written by the ACM SIGAI public policy officer) and AI Events (written by the ACM SIGAI conference coordination officer). The editors-in-chief have recently added an AI crossword puzzle (thanks to Adi Botea from IBM's Ireland Research Laboratory) and are about to add a column on current research trends in AI, written by recent grantees of research funds (such as NSF CAREER or European Research Council awards). AI Matters has also started to publish the winning student essays of the second ACM SIGAI Student Essay Contest.

ACM SIGAI also maintains an AI Matters blog (at `sigai.acm.org/aimatters/blog/`) as a forum for important announcements and news. For example, the ACM SIGAI public policy officer posts new information every two weeks in the blog to survey and report on current AI policy issues and raise awareness about the activities of other organizations that share interests with ACM SIGAI.

After a hiatus due to the illness of one of the organizers, ACM SIGAI recently restarted the ACM SIGAI webinars with a webinar on "Advances in Socio-Behavioral Computing" and several more in preparation. The webinars are streamed live but the videos can still be watched on demand at `learning.acm.org/webinar/`.

ACM SIGAI is also a founding member of AI Hub (at `aihub.org`), a new non-profit sibling to Robohub (at `robohub.org`) dedicated to connecting the AI communities of the world by bringing together experts in AI research, startups, business and education from across the globe. Content-area specialists will curate all incoming AI news articles to make sure that reporting is truthful, fair and balanced, and in-house editors will ensure that all content meets the highest editorial standards for language and clarity. AI Hub is expected to come online in Summer or Fall 2019. ACM SIGAI will provide content to AI Hub and, conversely, AI Hub will provide AI news to the ACM SIGAI members.

Financial Member Support

ACM SIGAI so far had concentrated its financial support on travel scholarships to ACM SIGAI student members to allow them to attend conferences if they are otherwise missing the financial resources to do so. The amounts of the scholarships vary but are generally in the range of US$1,000 to US$10,000 per conference, depending on the conference size. The ACM SIGAI conference coordination officer recently started to test a new open student award travel scheme. Beyond providing a ringfenced allocation to specific conferences, he created a process by which any ACM SIGAI student member who intends to attend an ACM (and, in exceptional cases, even a non-ACM) event can apply for travel support through the ACM SIGAI web-site. In the first few months since the inception of the scheme, students have already been offered financial support of about US$8,000 in total.

ACM SIGAI also recently created the AI Activities Fund, a new initiative to empower ACM SIGAI members and friends to organize activities with a strong outreach component to either students, researchers or practitioners not working on AI technologies or to the public in general. The purpose of the inaugural call for funding proposals was to help ACM SIGAI members and friends to promote a better understanding of current AI technologies, including their strengths and limitations as well as their promise for the future. Examples of fundable activities included (but were not limited to) AI technology exhibits or exhibitions, holding meetings with panels on AI technology (including on AI ethics) with expert speakers, creating podcasts or short films on AI technologies that are accessible to the public and holding AI programming competitions. ACM SIGAI was looking for evidence that the information presented by the activities would be of high quality, accurate, unbiased (for example, not influenced by company interests) and at the right level for the intended audience. The inaugural call for proposals supported the following initiatives: a workshop on "AI for All using the R Programming Language" organized by the Indian Institute of Technology in Goa, the "Bee Network of AI" organized by the Universidad Mayor in Chile and "Co-Opting AI: Public Conversations about Design, Inequality and Technology" organized by New York University.

Additional Member Services

ACM SIGAI also supports its members in additional ways. For example, it nominates them for awards or supports their nominations. ACM SIGAI is proud of the fact that many AI researchers in the past year received ACM honors, such as becoming ACM senior members, distinguished members and fellows as well as receiving other awards. Three AI researchers received the A.M. Turing Award in 2018.

ACM SIGAI also actively supports the Research Highlight Track of the Communications of the ACM (CACM) by nominating publications of recent, significant and exciting AI research results that are of interest to the computer science research community in general to the Research Highlight Track. This way, ACM SIGAI helps to make important AI research results visible to many computer scientists.

Additional ACM SIGAI membership benefits include reduced registration fees at many of the co-sponsored and in-cooperation conferences and access to the proceedings of many of these conferences in the ACM Digital Library.

Planning for the Future

ACM SIGAI held elections for a new chair, vice chair and secretary/treasurer in Spring 2019. Sanmay Das (the current ACM SIGAI vice chair) was elected ACM SIGAI chair, Nicholas Mattei (the current ACM SIGAI AI and society officer) was elected ACM SIGAI vice-chair, and John Dickerson (the current ACM SIGAI labor market officer) was elected ACM SIGAI secretary/treasurer. Sven Koenig (the current ACM SIGAI chair) will transition to his new role as ACM SIGAI past chair. We are looking forward to the new leadership committee shaping the future of ACM SIGAI. In general, ACM SIGAI intends to reach out to more AI groups worldwide that could benefit from ACM support, such as providing financial support, making the proceedings widely accessible in the ACM Digital Library and providing speakers via the ACM Distinguished Speakers program. ACM SIGAI also intends to reach out more to other disciplines that share an interest in AI, for example, in terms of research methodologies or applications.

Help Communities Solve Real-World Problems with AI – Become a Technovation Mentor!

Tara Chklovski (Founder and CEO, Technovation; Tara@technovation.org)

DOI: 10.1145/3362077.3362080

Abstract

Join other AI professionals as a mentor in Technovation's AI program for families. Share your expertise with adults and children who are curious about artificial intelligence and how it can be used to address real issues in their communities. Help people all around the world learn to not only use AI, but to create solutions with AI that improve their lives and their communities.

Technovation Families

Technovation, a global technology education nonprofit, is seeking mentors for its second season of Technovation Families, an AI-focused program for families with children ages 8-13. Join peers working in Computer Science fields and be part of the world's largest AI mentoring program. Support families around the world as they learn about AI and develop AI-based prototypes addressing problems they identify in their communities.

Started in 2018, Technovation Families' AI challenge brings families, schools, communities, and mentors together to learn, play, and create with AI. They apply what they learn to solve a real-world problem in their community as part of a global competition. Mentors guide learners of all ages through Neural Networks, data, self-driving car algorithms, and machine learning and training models to recognize images, text, and emotions through an IBM-Watson based platform Machine Learning for Kids. Local educators and CS experts offer encouragement and guidance throughout as families learn about – and use – AI tools for the first time. Mentors especially are a critical touch-point for families who are developing their confidence as problem solvers and inventors, and who have big ideas for applying AI to community problems, but lack technical knowledge, experience, and confidence in their abilities. Mentors are able to volunteer remotely, thereby strengthening local capacity in areas that may not have access to technology professionals or universities.

In the first year of Technovation Families' AI competition, 7,500 people across 70 chapters in 13 countries participated, developing 200 AI-based solutions to problems in their communities. These solutions ranged from image-recognition software that scans children's drawings for signs of bullying and a wearable swimming cap for kids to detect early signs of drowning, to a tool to detect and remove invasive algae from a local lake (Figure 1 and Figure 2).

Figure 1: Jeff Dean (Head of Google Brain) listening to a father and son team describing their image-recognition prototype that emits ultrasonic frequencies when it sees a dog.

NVIDIA - Tonie Hansen, Laura Clemmensen, Liz Austin
Patrick J. McGovern Foundation - Suzanne DiBianca, Patrick McGov
General Motors - Hin ch, Matthew Ybarra

Figure 2: Six coaches from Bolivia, Palestine, Spain, United States, Pakistan and Kazakhstan who coached their communities to create winning AI-based inventions.

Technovation partners with industry leaders including Google, NVIDIA, Intel, General Motors, and the Patrick J. McGovern Foundation, to bridge the AI knowledge and confidence gap for children and adults around the world (Figure 3).

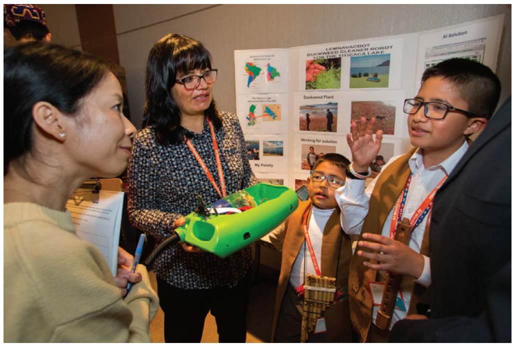
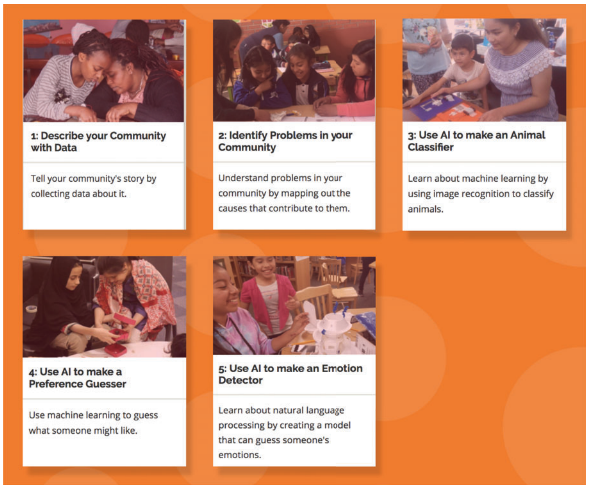


Figure 3: Mother and sons from Bolivia explaining to a judge how their Raspberry-Pi powered, image recognition system sucks up invasive weeds from Lake Titicaca.

The Technovation Families program is built on a community-based model that involves parents and caregivers so that the adults (in addition to the children) can reignite their curiosity and develop as lifelong learners. After participating in the program, more than 91% of the parents surveyed believed their child developed a sustained interest and growing interest in AI and ∼85% of parents wanted to continue investing effort into improving their local communities (Chklovski, Jung, Fofang, & Gonzales, 2019).

Mentors benefit too. Through Technovation's programs, mentors' communications and presentation skills improve, as do their professional relationships with colleagues and leaders. And, their work with Technovation participants stretches and grows their creativity, organizational, and project management skills.

Recently, the second season of Technovation Families launched debuting an expanded curriculum developed in partnership with a committee of AI researchers, and industry professionals. The updated curriculum includes additional information about AI, good and bad datasets, machine learning and ethical innovation (Figure 4 and Figure 5). Through 10 fun, hands-on lessons, families learn foundational AI concepts, identify a meaningful problem to solve in their local community, and build an AI agent to solve it. Sign-up today to help families make the world a better place with AI!



1: Describe your Community with Data

Tell your community's story by collecting data about it.

2: Identify Problems in your Community

Understand problems in your community by mapping out the causes that contribute to them.

3: Use AI to make an Animal Classifier

Learn about machine learning by using image recognition to classify animals.

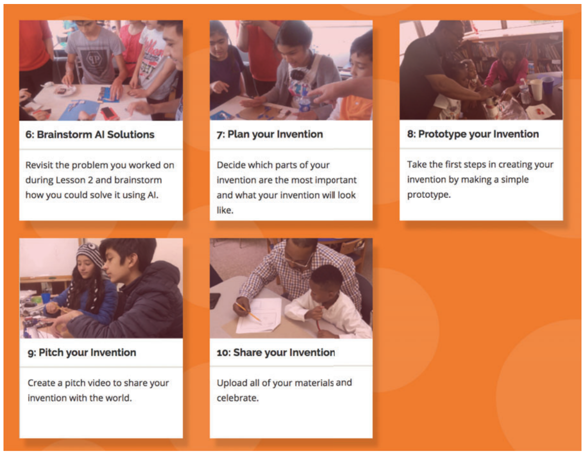
4: Use AI to make a Preference Guesser

Use machine learning to guess what someone might like.

5: Use AI to make an Emotion Detector

Learn about natural language processing by creating a model that can guess someone's emotions.

Figure 4: First 5-weeks of project-based Technovation Families AI curriculum that introduces learners to AI, Machine Learning, and building Image and Text Recognition Systems.



6: Brainstorm AI Solutions

Revisit the problem you worked on during Lesson 2 and brainstorm how you could solve it using AI.

7: Plan your Invention

Decide which parts of your invention are the most important and what your invention will look like.

8: Prototype your Invention

Take the first steps in creating your invention by making a simple prototype.

9: Pitch your Invention

Create a pitch video to share your invention with the world.

10: Share your Invention

Upload all of your materials and celebrate.

Figure 5: Last 5-weeks of the Technovation Families AI curriculum that helps participants apply their learning to create AI-based prototypes that address problems in their community.

References

Chklovski, T., Jung, R., Fofang, J. B., & Gonzales, P. (2019). Implementing a 15-week ai-education program with under-resourced families across 13 global communities. In *International joint conference on artificial intelligence, macau.*

Tara Chklovsk Tara Chklovski is CEO and founder of global tech education nonprofit Technovation. A frequent advocate for STEM education, she's presented at the White House STEM Inclusion Summit, SXSW EDU, UNESCO's Mobile Learning Week, and led the education track at the 2019 UN AI for Good Global Summit.

Events

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

DOI: 10.1145/3362077.3362081

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. For more information about conference support visit sigai.acm.org/activities/requesting_sponsorship.html.

2nd International Conference on Artificial Intelligence & Virtual Reality (AIVR 2019)

San Diego, CA, December 9-11, 2019
http://ieee-aivr.org
The AIVR conference, now in its second run, is a unique event, addressing researchers and industries from all areas of AI as well as Virtual, Augmented, and Mixed Reality. It provides an international forum for the exchange between those fields to present advances in the state of the art, identify emerging research topics, and together define the future of these exciting research domains. We invite researchers from VR, as well as Augmented Reality (AR) and Mixed Reality (MR) to participate and submit their work to the program. Likewise, any work on AI that has a relation to any of these fields or potential for the usage in any of them is welcome.

9th International Conference on Pattern Recognition Applications and Methods (ICPRAM '20)

Setúbal, Portugal, February 22-24, 2020
http://www.icpram.org/
The International Conference on Pattern Recognition Applications and Methods would like to become a major point of contact between researchers, engineers and practitioners on the areas of Pattern Recognition, both from theoretical and application perspectives. Contributions describing applications of Pattern Recognition techniques to real-world problems, interdisciplinary research, experimental and/or theoretical studies yielding new insights that advance Pattern Recognition methods are especially encouraged.
Submission deadline: October 4, 2019

13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020)

Valetta, Malta, February 24-26, 2020
http://www.biostec.org/
The purpose of BIOSTEC is to bring together researchers and practitioners, including engineers, biologists, health professionals and informatics/computer scientists, interested in both theoretical advances and applications of information systems, artificial intelligence, signal processing, electronics and other engineering tools in knowledge areas related to biology and medicine. BIOSTEC is composed of five co-located conferences, each specialized in a different knowledge area.
Submission deadline: October 4, 2019

12th International Conference on Agents and Artificial Intelligence (ICAART 2020)

Valetta, Malta, February 24-26, 2020
http://www.icaart.org/
The purpose of ICAART is to bring together researchers, engineers and practitioners interested in the theory and applications in the areas of Agents and Artificial Intelligence. Two simultaneous related tracks will be held, covering both applications and current research work. One track focuses on Agents, Multi-Agent Systems and Software Platforms, Distributed Problem Solving and Distributed AI in general. The other track focuses mainly on Artificial Intelligence, Knowledge Representation, Planning, Learning, Scheduling, Perception Reactive AI Systems, and Evolutionary Computing and other topics related to Intelligent Systems and Computational Intelligence.
Submission deadline: October 4, 2019

15th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2020)

Cambridge, UK, March 23-26, 2020
http://humanrobotinteraction.org/2020/
HRI 2020 is the 15th annual conference for basic and applied human-robot interaction research. Researchers from across the world present their best work to HRI to exchange ideas about the theory, technology, data, and science furthering the state-of-the-art in human-robot interaction. Each year, the HRI Conference highlights a particular area through a theme. The theme of HRI 2020 is "Real World Human-Robot Interaction". The HRI conference is a highly selective annual international conference that aims to showcase the very best interdisciplinary and multidisciplinary research in human-robot interaction with roots in and broad participation from communities that include but are not limited to robotics, artificial intelligence, human-computer interaction, human factors, design, and social and behavioral sciences. **Submission deadline: October 1, 2019**

22nd International Conference on Enterprise Information Systems (ICEIS 2020)

Prague, Czech Republic, May 5-7, 2020
http://www.iceis.org/
The purpose of ICEIS is to bring together researchers, engineers and practitioners interested in the advances and business applications of information systems. Six simultaneous tracks will be held, covering different aspects of Enterprise Information Systems Applications, including Enterprise Database Technology, Systems Integration, Artificial Intelligence, Decision Support Systems, Information Systems Analysis and Specification, Internet Computing, Electronic Commerce, Human Factors and Enterprise Architecture.
Submission deadline: December 13, 2019

19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2020)

Auckland, New Zealand, May 9-13, 2020
https://aamas2020.conference.auckland.ac.nz/
AAMAS is the leading scientific conference for research in autonomous agents and multi-agent systems. The AAMAS conference series was initiated in 2002 as the merging of three respected scientific meetings: the International Conference on Multi-Agent Systems (ICMAS), the International Workshop on Agent Theories, Architectures, and Languages (ATAL), and the International Conference on Autonomous Agents (AA). The aim of the joint conference is to provide a single, high-profile, internationally-respected archival forum for scientific research in the theory and practice of autonomous agents and multi-agent systems. AAMAS 2020 is the 19th edition of the AAMAS conference, and the first time AAMAS will be held in New Zealand. The conference solicits papers addressing original research on autonomous agents and their interaction, including agents that interact with humans. In addition to the main track, there will be two special tracks: Blue Sky Ideas and JAAMAS.
Submission deadline: November 15, 2020

33rd International Conference on Industrial, Engineering and Other Applications (IEA/AIE '20)

Kitakyushu, Japan, July 21-24, 2020
https://jsasaki3.wixsite.com/ieaaie2020
IEA/AIE 2020 continues the tradition of emphasizing on applications of applied intelligent systems to solve real-life problems in all areas including engineering, science, industry, automation & robotics, business & finance, medicine and biomedicine, bioinformatics, cyberspace, and human-machine interactions.
Submission deadline: December 15, 2019

35th IEEE/ACM International Conference on Automated Software Engineering (ASE 2020)

Melbourne, Australia, September 21-25, 2020
https://www.deakin.edu.au/ase2020
The 35th IEEE/ACM International Conference on Automated Software Engineering (ASE 2020) will be held in Melbourne, Australia from September 21 to 25, 2020. The conference is the premier research forum for automated software engineering. Each year, it brings together researchers and practitioners from academia and industry to discuss foundations, techniques, and tools for automating

the analysis, design, implementation, testing, and maintenance of large software systems.

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

AI Education Matters: Building a Fake News Detector

Michael Guerzhoy (Princeton University, University of Toronto, and the Li Ka Shing Knowledge Institute, St. Michael's Hospital; guerzhoy@princeton.edu)

Lisa Zhang (University of Toronto Mississauga; lczhang@cs.toronto.edu)

Georgy Noarov (Princeton University; gnoarov@princeton.edu)

DOI: 10.1145/3362077.3362082

Introduction

Fake news is a salient societal issue, the subject of much recent academic research, and, as of 2019, a ubiquitous catchphrase.

In this article, we explore using the task of detecting fake news to teach supervised machine learning and data science, as demonstrated in our Model AI Assignment[1] (Neller et al., 2019). We ask students to build a series of increasingly complex classifiers that categorize news headlines into "fake" and "real" and to analyze the classifiers they have built. Students think about the data, the validity of the problem posed to them, and the assumptions behind the models they use. Students can compete in a class-wide competition to build the best fake news detector.

To help instructors incorporate fake news detection into their course, we briefly review recent research on fake news and the task of fake news detection. We then describe the assignment design, and reflect on the in-class fake news detection competition we ran.

Fake News Research

Fake news is an old issue (Mansky, 2018), but the role it may have played in the 2016 US Presidential Election has sparked renewed interest in the phenomenon (Lazer et al., 2018), (Allcott & Gentzkow, 2017). Research on fake news is focused on understanding its audience and societal impact, how it spreads on social media, and who its consumers are (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019), (Nelson & Taneja, 2018).

Fake news can be detected based on textual features and social network propagation patterns (Shu, Sliva, Wang, Tang, & Liu, 2017). High-quality datasets of fake and real news are scarce. Several medium-scale datasets have recently been collected, with fake news either obtained from the web (often with the help of fact-checking resources such as `PolitiFact.com`) or written to order by Amazon Mechanical Turk workers (Wang, 2017), (Pérez-Rosas, Kleinberg, Lefevre, & Mihalcea, 2018).

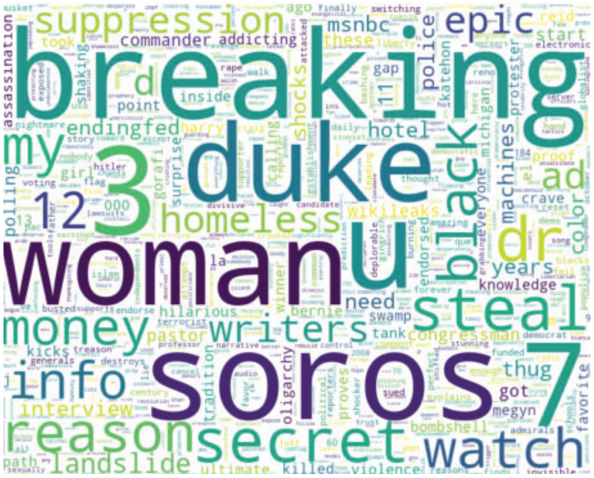


Figure 1: Visualizing $P(\text{fake}|\text{keyword})$ for a naive Bayes model trained on our training set. Larger text corresponds to larger conditional probabilities.

The definition of the concept of "fake news" itself has proven elusive. See (Tandoc, Lim, & Ling, 2018) for an overview of the definitions recently used in literature.

Teaching Supervised Learning via Fake News

In our assignment, the task is to classify news headlines as "real" or "fake". Students build and compare several standard classifiers: naive Bayes, logistic regression, and a decision tree. All three classifiers use the

[1] `http://modelai.gettysburg.edu/2019/fakenews/`

presence/absence of keywords as their feature set. The detection of fake news headlines using naive Bayes is directly analogous to the classic spam filtering task. See (Russell & Norvig, 2009) for an exposition and (Sahami, Dumais, Heckerman, & Horvitz, 1998) for the paper that introduced the idea.

Our pedagogical approach emphasizes having students analyze the models they build. In particular, we ask students to obtain keywords whose presence or absence most strongly indicates that a headline is "fake" or "real".

To find the most important keywords for classifying a headline as "real" using naive Bayes, students need to decide whether they should use $P(\text{real}|\text{keyword})$ or $P(\text{keyword}|\text{real})$. We hope they gain a deeper understanding of naive Bayes in the process. We use PyTorch to implement logistic regression, and suggest (as would be natural for our students) that students use multinomial logistic regression with 2 outputs when predicting "fake"/"real". This results in $2k + 2$ coefficients for a vocabulary of k keywords. Identifying the most important keywords based on these $2k + 2$ numbers nudges students towards understanding the details of the model. We also ask students to derive the logistic regression coefficients that correspond to the naive Bayes classifier they fit. As a final step, students fit a decision tree to the data and again identify the most important features according to the model.

As they fit a series of increasingly complex classifiers, students observe overfitting firsthand: training performance increases with classifier complexity, while validation performance decreases. Beating naive Bayes turns out to be quite difficult (though doable). Students attempt to do that in the competition phase.

Teaching Data Science via Fake News

When using the assignment in a data science rather than a machine learning course, we place more emphasis on statistical modeling and careful examination of the data. We ask students to inspect the dataset in order to analyze it qualitatively and discuss its limitations. Students are also asked to check whether the dataset conforms to the naive Bayes assumption (it does not; to figure out why, students need to think about how human language works).

Another part of the assignment involves producing new data via the naive Bayes generative model. The goal is for students to gain a deeper understanding of generative models.

Datasets

The dataset students use in the principal part of the assignment was compiled by combining data from several sources. It consists of 1298 "fake news" headlines and 1968 "real news" headlines, all containing the word "Trump". "Fake" headlines are challenging to collect; as students see, most headlines labeled as such could be argued to be merely tendentious or hyperbolic rather than fake.

We have curated a smaller private test set of headlines that we have verified to be either real or fake[2]. That test set is used in our fake news detection competition and is available to instructors upon request.

Fake News Detection Contest

For interested students, we ran an optional fake news detection competition. The authors of the best-performing entries would earn a small amount of points towards their course grade. Participating students could follow any approach they liked. We encouraged augmenting the given training set with more data, engineering useful features, training classifiers of the students' own choice, and using ensemble methods. Gratifyingly, some contestants were able to engineer useful features and use modern text classification algorithms to beat the naive Bayes baseline.

The source code for many modern text classification systems is widely available and sometimes comes with pre-trained weights. Students would often adapt, train, or fine-tune the systems for their submissions.[3]

[2] While we could not fact-check the headlines to journalistic standards, we made sure that the truth or falseness of the headlines was not in serious dispute.

[3] Training deep learning systems is often resource intensive. We refer students to services such as AWS, Microsoft Azure, and Google Cloud Platform, where they are eligible for free credits.

Conclusion

Through building a fake news detector in class, we are able to teach some of the foundational methods of supervised learning in a compelling and coherent manner. The dataset we collected can be used in a class that emphasizes rigorous thinking about data science problems. We share our experience of running an in-class fake news detection competition.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–36.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, *363*(6425), 374–378.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... others (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Mansky, J. (2018). The age-old problem of "fake news". *Smithsonian Magazine*. https://www.smithsonianmag.com/history/age-old-problem-fake-news-180968945/.

Neller, T. W., Sooriamurthi, R., Guerzhoy, M., Zhang, L., Talaga, P., Archibald, C., ... others (2019). Model AI Assignments 2019. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 9751–9753).

Nelson, J. L., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New Media & Society*, *20*(10), 3720-3737.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3391–3401).

Russell, S. J., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Pearson Education Limited.

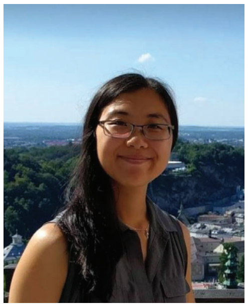
Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop* (Vol. 62, pp. 98–105).

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news". *Digital Journalism*, *6*(2), 137-153.

Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.*

Michael Guerzhoy is a Lecturer at Princeton University, an Assistant Professor (Status Only) at the University of Toronto, and a Scientist at the Li Ka Shing Knowledge Institute, St. Michael's Hospital. His professional interests are in computer science and data science education and in applications of machine learning to healthcare.



Lisa Zhang is an Assistant Professor, Teaching Stream (CLTA) at the University of Toronto Mississauga. Her current research interests are in the intersection of computer science education and machine learning.

Georgy Noarov is a student at Princeton University concentrating in mathematics and pursuing a certificate in statistics and machine learning. His research areas include algorithmic game theory and combinatorial optimization.

AI Education Matters: A First Introduction to Modeling and Learning using the Data Science Workflow

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

DOI: 10.1145/3362077.3362083

Introduction

Traditionally artificial intelligence (AI) and machine learning (ML) courses are taught at the senior and graduate level in higher-education computer science curricula following the *mastery learning* strategy, cf. Figure 1. This makes sense, since most AI and ML models and the theory behind them require a substantial understanding of probability and statistics, as well as advanced calculus and matrix algebra. To understand Logistic Regression as a probabilistic classifier performing maximum-likelihood or maximum-a-posteriori estimation, for example, students need to understand joint and conditional probability distributions. In order to derive the back propagation algorithm to train Neural Networks students need to understand partial derivatives and inner and outer tensor products. These are just two of many examples where substantial mathematical background – typically taught at the junior level in a computer science major program – is required. With AI and ML algorithms being used more widely by enterprises across domains, as well as, in applications and services we use in our daily lives, it makes sense to raise awareness about what AI is, what it can and cannot do, and how it is used to solve problems to a broader audience. Very much in the same spirit as the "CS for all" idea (https://www.csforall.org), we have to extend our curricula to include introductory courses to AI and ML on the early undergraduate level (or even in highschool) to expose students to the ideas and working principles of AI technology. One way to achieve this is to introduce the principles of working with data, modeling, and learning through the data science workflow.

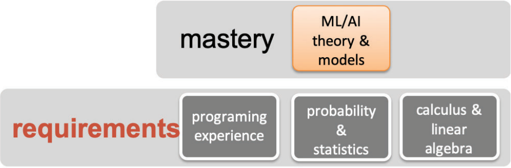


Figure 1: Mastery learning paradigm.

Exposure First

Following the *exposure – interest – mastery* paradigm as illustrated in Figure 2, we propose to gently introduce AI/ML concepts focusing on example applications rather than computational problems by incorporating course modules into introductory CS courses or design an entire course early on the curriculum. The goal of such intro-level modules or courses is to expose students to AI/ML problems and introduce basic techniques to solve them without relying on the computational and mathematical prerequisite knowledge. More concretely, the module or course may be designed as combined lecture and lab sessions, where a new topic is introduced in a lecture unit followed by a lab session, where students get to know a problem in the context of an application, explore a solution method, and tackle a potentially open-ended question about evaluation procedures, benefits and challenges of the approach, or implications and ethical considerations when using such methods in the real world in a group discussion. Lab sessions should be designed carefully focusing on the understanding of the data, the problem, and the results instead of model implementation. We will introduce two such lab assignments implemented in Python

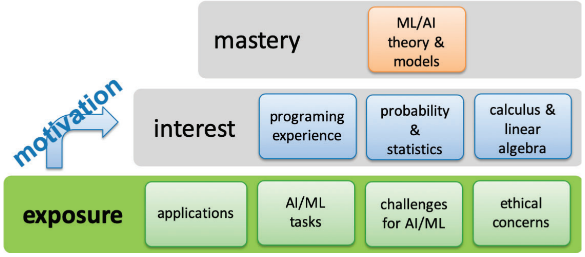


Figure 2: Exposure-Interest-Mastery paradigm.

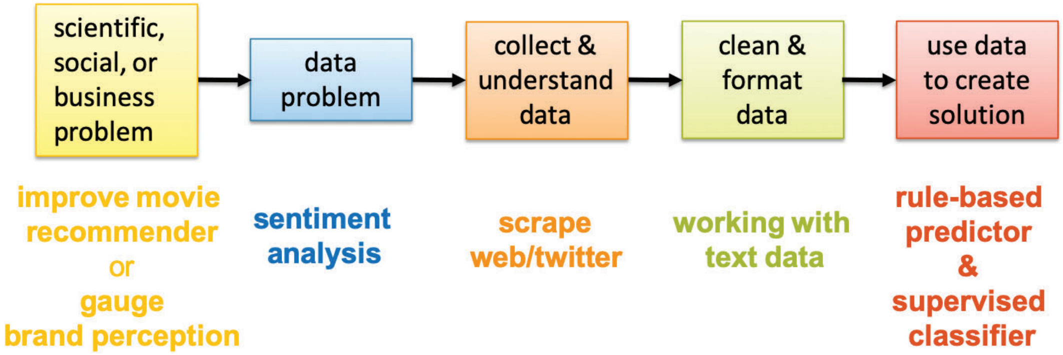

scientific, social, or business problem
data problem
collect & understand data
clean & format data
use data to create solution
improve movie recommender or gauge brand perception
sentiment analysis
scrape web/twitter
working with text data
rule-based predictor & supervised classifier

Figure 3: Data science workflow using sentiment analysis as an example application.

and Jupyter notebooks in the next section.

The main aim of our assignments is to engage the students' interest to acquire the prerequisite knowledge in order to move forward and gain a deeper understating of specific AI and ML techniques. Since we propose these units for a course that is taught very early in the CS curriculum, we face the challenge that students do not have a lot of programming experience nor a deep understanding of data structures and algorithms. Therefore, we developed the lab assignments using Jupyter notebooks which nicely combine illustrative instructions and executable starter code.

After having worked though the data science workflow using illustrative applications that are both easy to understand and relevant in the real-world, our hope is that students develop the motivation to study traditional prerequisite classes for AI and ML courses like probability and statistics, matrix/linear algebra, and algorithm analysis perceiving them useful to master AI/ML instead of a nuisance.

Two Model AI Assignments

Introduction to Python for Data Science

We provide an interactive guided lab to introduce Python for data science (DS),[1] which can also be used for any course that introduces modeling and learning using Python, such as introduction to AI or ML courses. We provide two Jupyter notebooks, one introducing the basics of Python and the other the DS workflow using the Iris dataset (https://archive.ics.uci.edu/ml/datasets/Iris). We interactively introduce the use of expressions, variables, strings, printing, lists, dictionaries, control flow, and functions in Python to students that are already familiar with a programming language from an introductory CS course. The second lab aims at motivating students to acquire skills such as using statistics to model and analyze data, knowing how to design and use algorithms to store, process, and visualize data, while not forgetting the importance of domain expertise. We begin by establishing the example problem to be studied based on the Iris dataset. The next step is to acquire and process the data, where students practice how to load data and process strings into numeric arrays using `numpy`. Then, we explain different plotting methods such as box plots, histograms, and scatter plots for data exploration leveraging `matplotlib`. Finally, we split the data into training and test set, build a model, use it for predictions, and evaluate the results using `sklearn`. The main learning objectives are to get to know and practice Python in the context of a realistic data science and machine learning application.

Introducing the Data Science Workflow using Sentiment Analysis

The second interactive lab guides students through a basic data science workflow by exploring sentiment analysis.[2] The data science workflow along with the example sentiment analysis application is depicted in Figure 3. The lab assignment focuses on introducing

[1] http://modelai.gettysburg.edu/2019/intro2py/

[2] http://modelai.gettysburg.edu/2019/intro2ds/

the machinery using a given dataset of movie reviews. We further provide a follow-up homework assignment reiterating some of the steps and highlighting data acquisition and exploration with Twitter data. After introducing sentiment analysis, we explain a simple rule-based approach to predict the sentiment of textual reviews using three handcrafted examples. This introduction shows simple means to preprocess text data and exemplifies the use of lists of positive and negative expressions to compute a sentiment score. Then students will implement the approach to predict the sentiment of movie reviews and evaluate the results. The lab concludes with a discussion of the limitations of the rule-based approach and a quick introduction to sentiment classification via machine learning. The homework assignment reiterates over the process of building and analyzing a sentiment predictor with the focus on collecting and preprocessing their own dataset scraped from Twitter using the `python-twitter` API. The main learning objective of this activity is getting to know the inference problem and walking through the entire data science workflow to tackle it. Since the module only requires minimal programming background it is an ideal precursor to introducing machine learning in an AI, ML, or DS course. It may also be used in a introduction to Python course as a module focusing on using libraries and APIs.

Our Experiences

We incorporated both lab assignments into our “Introduction to Data Science” course for sophomore students at Washington University in St. Louis. One of the challenges we faced was that our students had different levels of Python experience, from no experience at all (51%) over some experience (36%) to quite proficient (13%). This led to a large variance in the times needed to complete the labs. To deal with this issue we propose to add some optional challenge problems to the assignment that are not required for the homework or will be introduced later in the course. Another challenge was that some students preferred to work in groups where others did the labs on their own. However, both strategies can result in slower or faster pace given the students’ working style, group composition, and amount of group discussion. Unfortunately, there is no unified way to tackle this issue, however, we believe that students should be encouraged to work in teams for the lab assignments, whereas homework assignments should be worked on individually. This way both teamwork and communication skills as well as knowledge retention are facilitated.

Both labs were perceived as useful by our students. 97% answered *Yes* to the question “Did you like the lab.” for the introduction to Python lab and 81% for the sentiment analysis lab. The most common reasons stated by students that didn’t like the second lab were that they where overwhelmed by unfamiliar code and that it was too long. From the students’ answers to our quiz and exam questions we can also confirm that they understand basic Python processes to handle data, implement and apply simple learning models, and visualize and interpret their results.

Pedagogical Resources

In addition to Jupyter notebooks constituting the lab and homework assignments, we developed lecture materials in form of slides and worksheets for each module. The first lecture covers an introduction to data science and machine learning, and the second one introduces sentiment analysis, text processing, and classification respectively. The slides are interactive with gaps to be filled in by the instructor during the lectures and the worksheets contain in-class activities for students to engage with the presented materials. Those resources are available from the authors upon request.

Useful textbooks that specifically focus on introducing data science topics and techniques are:

- Python Data Science Handbook VanderPlas (2016) introduces essential tools and libraries such as Jupyter notebooks, numpy, pandas, scikit-learn, and matplotlib for working with data.
- Data Science from Scratch Grus (2019) focuses on implementing learning algorithms and data processing routines from scratch.
- Data Science for Business Provost and Fawcett (2013) showcases interesting real-world use cases and emphasizes data-

analytic thinking while not being too technical.

The first two books focus on implementations in Python, whereas the third one details concepts and techniques without code examples.

References

Grus, J. (2019). *Data science from scratch: first principles with python*. O'Reilly Media.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.".

VanderPlas, J. (2016). *Python data science handbook: essential tools for working with data*. " O'Reilly Media, Inc.".

Marion Neumann is a Senior Lecturer at Washington University in St. Louis and the SIGAI diversity officer. She teaches Machine Learning, Cloud Computing, Analysis of Networked Data, and Introduction to Data Science. Her research interests include graph-based machine learning and analyzing networked data as well as measuring and analyzing student emotions in large computing courses using sentiment analysis.

AI Policy Matters

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

DOI: 10.1145/3362077.3362084

Abstract

AI Policy Matters 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.

About Face

Face recognition (FR) research has made great progress in recent years and has been prominent in the news. In public policy, many are calling for a reversal of the trajectory for FR systems and products. In the hands of people of good will, using products designed for safety and training systems with appropriate data, FR benefits society and individuals. *The Verge* reports the use in China of unique facial markings of pandas to identify individual animals. FR research includes work to mitigate negative outcomes, as with the Adobe and UC Berkeley work on Detecting Facial Manipulations in Adobe Photoshop for automatic detection of facial images that have been manipulated by splicing, cloning, and removing objects.

Intentional and unintentional application of systems that are not designed and trained for ethical use are a threat to society. Screening for terrorists could be good, but FR lie and fraud detection systems sometimes do not work properly. The safety of FR is currently an important issue for policymakers, but regulations could have negative consequences for AI researchers. As with many contemporary issues, conflicts arise because of conflicting policies in different countries. Recent and current legislation is attempting to restrict FR use and possibly inhibit FR research; for example,

- San Francisco, CA, Somerville, MA, and Oakland, CA, are the first three cities to limit use of FR to identify people.
- In “Facial recognition may be banned from public housing thanks to proposed law” CNET reports that a bill will be introduced to address the issue that “landlords across the country continue to install smart home technology and tenants worry about unchecked surveillance.”
- A call for a more comprehensive ban on FR has been launched by the digital rights group Fight for the Future, seeking a complete Federal ban on government use of facial recognition surveillance.

Beyond legislation against FR research and banning certain products, work is in progress to enable safe and ethical use of FR. A more general example that could be applied to FR is the MITRE work The Ethical Framework for the Use of Consumer-Generated Data in Health Care, which “establishes ethical values, principles, and guidelines.”

AI Regulation

With AI in the news so much over the past year, the public awareness of potential problems arising from the proliferation of AI systems and products has led to increasing calls for regulation. The popular media, and even technical media, do contain misinformation and misplaced fears, but plenty of legitimate issues exist even if their relative importance is sometimes misunderstood. Policymakers, researchers, and developers need to be in dialog about the true needs for and potential dangers of regulation. From our policy perspective, the significant risks from AI systems include misuse and faulty unsafe designs that can create bias, non-transparency of use, and loss of privacy. Some AI systems are known to discriminate against minorities, unintentionally and not.

An important discussion we should be having is if governments, international organizations, and big corporations, which have already released dozens of non-binding guidelines for the responsible development and use of AI, are the best entities for writing and enforcing regulations. Non-binding principles will not make some companies developing and applying AI products accountable. An important point in this regard is to hold companies responsible for the product design process itself, not just for testing products after they are in use.

Introduction of new government regulations is a long process and subject to pressure from lobbyists, and the current US administration is generally inclined against regulations anyway. We should discuss alternatives like clearinghouses and consumer groups endorsing AI products designed for safety and ethical use. If well publicized, the endorsements of respected non-partisan groups including professional societies might be more effective and timely than government regulations.

The European Union has released its Ethics Guidelines for Trustworthy AI, and a second document with recommendations on how to boost investment in Europe's AI industry is to be published. In May, 2019, the Organization for Economic Cooperation and Development (OECD) issued their first set of international OECD Principles on Artificial Intelligence, which are embraced by the United State and leading AI companies.

The AI Race

China, the European Union, and the United States have been in the news about strategic plans and policies on the future of AI. The U.S. National Artificial Intelligence Research and Development Strategic Plan, was released in June, 2019, as an update of the report by the Select Committee on Artificial Intelligence of The National Science and Technology Council. The Computing Community Consortium (CCC) recently released the AI Roadmap Website.

Now, the Center for Data Innovation has issued a Report comparing the current standings of China, the European Union, and the United States. Here is a summary of their policy recommendations: "Many nations are racing to achieve a global innovation advantage in artificial intelligence (AI) because they understand that AI is a foundational technology that can boost competitiveness, increase productivity, protect national security, and help solve societal challenges. This report compares China, the European Union, and the United States in terms of their relative standing in the AI economy by examining six categories of metrics: talent, research, development, adoption, data, and hardware. It finds that despite the bold AI initiatives in China, the United States still leads in absolute terms. China comes in second, and the European Union lags further behind. This order could change in coming years as China appears to be making more rapid progress than either the United States or the European Union. Nonetheless, when controlling for the size of the labor force in the three regions, the current U.S. lead becomes even larger, while China drops to third place, behind the European Union. This report also offers a range of policy recommendations to help each nation or region improve its AI capabilities."

US and G20 AI Policy

The G20 AI Ministers from the Group of 20 major economies conducted meetings on trade and the digital economy. They produced guiding principles for using artificial intelligence based on principles adopted earlier by the 36-member OECD and an additional six countries. The G20 guidelines call for users and developers of AI to be fair and accountable, with transparent decision-making processes and to respect the rule of law and values including privacy, equality, diversity and internationally recognized labor rights. Meanwhile, the principles also urge governments to ensure a fair transition for workers through training programs and access to new job opportunities.

Bipartisan Legislators On Deepfake Videos

Senators introduced legislation intended to lessen the threat posed by "deepfake" videos, which use AI technologies to manipulate original videos and produce misleading information. With this legislation, the Department of Homeland Security would conduct an annual study of deepfakes and related content and require the department to assess the AI technologies used to create deepfakes. This could lead to changes in regulations or to new regulations impacting the use of AI.

Hearing on Societal and Ethical Impacts

The House Science, Space and Technology Committee held a hearing on June 26th about the societal and ethical implications of artificial intelligence, now available on video. The National Artificial Intelligence Research and Development Strategic Plan, released in June, is an update of the report by the Select Committee on Artificial Intelligence of The National Science and Technology Council.

On February 11, 2019, the President signed Executive Order 13859: Maintaining American Leadership in Artificial Intelligence. According to Michael Kratsios, Deputy Assistant to the President for Technology Policy, this order "launched the American AI Initiative, which is a concerted effort to promote and protect AI technology and innovation in the United States. The Initiative implements a whole-of-government strategy in collaboration and engagement with the private sector, academia, the public, and like-minded international partners. Among other actions, key directives in the Initiative call for Federal agencies to prioritize AI research and development investments, enhance access to high-quality cyberinfrastructure and data, ensure that the Nation leads in the development of technical standards for AI, and provide education and training opportunities to prepare the American workforce for the new era of AI."

The first seven strategies continue from the 2016 Plan, reflecting the reaffirmation of the importance of these strategies by multiple respondents from the public and government, with no calls to remove any of the strategies. The eighth strategy is new and focuses on the increasing importance of effective partnerships between the Federal Government and academia, industry, other non-Federal entities, and international allies to generate technological breakthroughs in AI and to rapidly transition those breakthroughs into capabilities.

Strategy 8: Expand Public–Private Partnerships to Accelerate Advances in AI is new in the June, 2019, plan and reflects the growing importance of public-private partnerships enabling AI research and expanding public-private partnerships to accelerate advances in AI. A goal is to promote opportunities for sustained investment in AI research and development and transitions into practical capabilities, in collaboration with academia, industry, international partners, and other non-Federal entities.

Continued points from the seven Strategies in the previous Executive Order in February include

- support for the development of instructional materials and teacher professional development in computer science at all levels, with emphasis at the K–12 levels,
- consideration of AI as a priority area within existing Federal fellowship and service programs,
- development of AI techniques for human augmentation,
- emphasis on achieving trust: AI system designers need to create accurate, reliable systems with informative, user-friendly interfaces.

The National Science and Technology Council (NSTC) is functioning again. NSTC is the principal means by which the Executive Branch coordinates science and technology policy across the diverse entities that make up the Federal research and development enterprise. A primary objective of the NSTC is to ensure that science and technology policy decisions and programs are consistent with the President's stated goals. The NSTC prepares research and development strategies that are coordinated across Federal agencies aimed at accomplishing multiple national goals. The work of the NSTC is organized under committees that oversee subcommittees and working groups focused on different aspects of science and technology. More information is available.

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, homeland security, health, foreign relations, the environment, and the technological recovery and use of resources, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget with an annual review and analysis of Federal research and development in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the Federal Government. More information is available.

Groups that advise and assist the NSTC on AI include

- The Select Committee on Artificial Intelligence addresses Federal AI research and development activities, including those related to autonomous systems, biometric identification, computer vision, human computer interactions, machine learning, natural language processing, and robotics. The committee supports policy on technical, national AI workforce issues

- The Subcommittee on Machine Learning and Artificial Intelligence monitors the state of the art in machine learning (ML) and artificial intelligence within the Federal Government, in the private sector, and internationally

- The Artificial Intelligence Research and Development Interagency Working Group coordinates Federal research and development in AI and supports and coordinates activities tasked by the Select Committee on AI and the NSTC Subcommittee on Machine Learning and Artificial Intelligence.

More information available.

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.

Advancing Non-Convex and Constrained Learning: Challenges and Opportunities

Tianbao Yang (The University of Iowa; tianbao-yang@uiowa.edu)

DOI: 10.1145/3362077.3362085

Introduction

As data gets more complex and applications of machine learning (ML) algorithms for decision-making broaden and diversify, traditional ML methods by minimizing an unconstrained or simply constrained convex objective are becoming increasingly unsatisfactory. To address this new challenge, recent ML research has sparked a *paradigm shift* in learning predictive models into non-convex learning and heavily constrained learning. Non-Convex Learning (NCL) refers to a family of learning methods that involve optimizing non-convex objectives. Heavily Constrained Learning (HCL) refers to a family of learning methods that involve constraints that are much more complicated than a simple norm constraint (e.g., data-dependent functional constraints, non-convex constraints), as in conventional learning. This paradigm shift has already created many promising outcomes: (i) non-convex deep learning has brought breakthroughs for learning representations from *large-scale structured data* (e.g., images, speech) (LeCun, Bengio, & Hinton, 2015; Krizhevsky, Sutskever, & Hinton, 2012; Amodei et al., 2016; Deng & Liu, 2018); (ii) non-convex regularizers (e.g., for enforcing sparsity or low-rank) could be more effective than their convex counterparts for learning *high-dimensional structured models* (C.-H. Zhang & Zhang, 2012; J. Fan & Li, 2001; C.-H. Zhang, 2010; T. Zhang, 2010); (iii) constrained learning is being used to learn predictive models that satisfy various constraints to *respect social norms* (e.g., fairness) (B. E. Woodworth, Gunasekar, Ohannessian, & Srebro, 2017; Hardt, Price, Srebro, et al., 2016; Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017; A. Agarwal, Beygelzimer, Dudík, Langford, & Wallach, 2018), to *improve the interpretability* (Gupta et al., 2016; Canini, Cotter, Gupta, Fard, & Pfeifer, 2016; You, Ding, Canini, Pfeifer, & Gupta, 2017), to *enhance the robustness* (Globerson & Roweis, 2006a; Sra, Nowozin, & Wright, 2011; T. Yang, Mahdavi, Jin, Zhang, & Zhou, 2012), etc. In spite of great promises brought by these new learning paradigms, they also bring emerging challenges to the design of computationally efficient algorithms for *big data* and the analysis of their statistical properties.

Non-Convex Learning

In this section, we describe some recent advances in non-convex learning with mentioning some of our recent related results. We will also describe their limitations and point out future directions. This article will focus on studies that are concerned with algorithm design and analysis for solving NCL and HCL problems instead of papers that are purely application-driven. It is notable that the references are not exhaustive due to a large volume of related works.

Non-Convex Minimization and Deep Learning. Deep learning can be formulated as the following non-convex minimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) := \mathrm{E}_{\mathbf{z}}[f(\mathbf{w};\mathbf{z})], \qquad (1)$$

where $\mathbf{z}$ denotes a random data, and $\mathbf{w}$ denotes the parameters of the neural network to be learned, and $f(\mathbf{w};\mathbf{z})$ denotes the loss function. Due to the success of deep learning in many areas, this problem has attracted much attention from the community of mathematical programming and machine learning. Research has been conducted in the following directions.

- **Convergence to stationary points.** For general non-convex problems, it is NP-hard to find a global minimizer (Hillar & Lim, 2013). Hence, many studies have focused on finding stationary points of (1) (Nesterov & Polyak, 2006; N. Agarwal, Allen Zhu, Bullins, Hazan, & Ma, 2017; Carmon, Duchi, Hinder, & Sidford, 2016; P. Xu, Roosta-Khorasani, & Mahoney, 2017; Cartis, Gould, & Toint, 2011b, 2011a; Royer & Wright, 2017; M. Liu & Yang, 2017b, 2017a; Allen-Zhu, 2017; Kohler & Lucchi, 2017; Reddi et al., 2017). Typically, two types of stationary points are considered, namely first-order stationary point and second-order stationary point. A point $\mathbf{w}_*$ is called a first-order stationary point if it satisfies $\nabla F(\mathbf{w}_*) = 0$. A point $\mathbf{w}_*$ is called a second-order stationary if it satisfies $\nabla F(\mathbf{w}_*) = 0$ and $\nabla^2 F(\mathbf{w}_*) \succeq 0$. These studies concentrate on the complexity analysis of first or second-order methods. Many first-order methods (e.g., stochastic gradient descent (SGD)) have been proved to converge to first-order stationary points with a polynomial time complexity. In our study (Yan, Yang, Li, Lin, & Yang, 2018), we presented the first theoretical result showing that the commonly used stochastic heavy-ball (SHB) method and stochastic Nesterov's accelerated gradient (SNAG) method for deep learning converge to first-order stationary points, and also presented a unified framework that subsumes SGD, SHB and SNAG by varying a single parameter. Moreover, in (Y. Xu, Rong, & Yang, 2018) we presented a unified framework that can promote first-order algorithms to enjoy convergence to a second-order stationary point by using our proposed first-order negative curvature finding procedure named NEON.

- **Convergence to global minimizers.** Recently, several works have proved gradient descent or stochastic gradient descent can find global minimizers of minimizing an over-parameterized deep neural network under some mild conditions of input data (Allen-Zhu, Li, & Song, 2018; Arora, Cohen, & Hazan, 2018; Y. Li & Liang, 2018; Du, Zhai, Poczos, & Singh, 2018; Zou, Cao, Zhou, & Gu, 2018). Different from other studies that focus on general non-convex minimization problems, these recent works explored the properties for overparameterized deep neural networks and presented sharp analysis of (stochastic) gradient descent.

- **Smart Step Sizes or Learning Rates.** Step sizes or learning rates play an important role in an optimization algorithm for learning deep neural networks. Conventional polynomially decreasing step sizes are observed to be non-effective for deep learning. Smart step size schemes have been proposed including stagewise geometrically decreasing step size (Y. Xu, Lin, & Yang, 2017), and adaptive step sizes (Kingma & Ba, 2015; J. Chen & Gu, 2018; Zhou, Tang, Yang, Cao, & Gu, 2018; Zaheer, Reddi, Sachan, Kale, & Kumar, 2018; Luo, Xiong, Liu, & Sun, 2019; Z. Chen et al., 2019). A stagewise geometrically decreasing step size is usually adopted in SGD, SHB and SNAG for deep learning, which starts from a relatively large step size and decreases by a constant factor after a number of iterations. This step size scheme has achieved the state of the art result on the ImageNet classification task (He, Zhang, Ren, & Sun, 2016; Real, Aggarwal, Huang, & Le, 2019; Tan & Le, 2019). The idea of adaptive step size dates back to AdaGrad (Duchi, Hazan, & Singer, 2011), which was proposed for convex optimization. It has several variants with Adam (Kingma & Ba, 2015) being one of its most popular variants. The adaptive algorithms have been analyzed for non-convex optimization problems (X. Li & Orabona, 2018; J. Chen & Gu, 2018; Zhou et al., 2018; Zaheer et al., 2018; Luo et al., 2019).

Limitations and Future Directions. Although some nice results have been achieved in non-convex optimization and learning deep neural networks, there still remain many issues that require further investigation.

- **The gap between practice and theory.** There are several limitations of existing analysis: (i) most existing analysis of SGD uses a very small step size (Ghadimi & Lan, 2013; Yan et al., 2018; Davis & Drusvyatskiy, 2018), which is far from being practical; (ii) most theoretical analysis of non-convex optimization algorithms focus on optimization error; however, it is more important to consider the generalization performance of a stochastic optimization algorithm; (iii) global analysis of SGD imposes strong conditions on the level of over-parameterization (Allen-Zhu et al., 2018; Arora et al., 2018; Y. Li & Liang, 2018; Du et al., 2018; Zou et al., 2018), which is far from being practical. To address the first two limitations, we have conducted some preliminary study of SGD with a stagewise geometrically decreasing step size scheme by analyzing both the optimization error and the generalization error. Our analysis exhibits that the stagewise geometrically decreasing step scheme can leverage some nice properties of deep neural networks and enjoy faster convergence for both the training error and testing error than using a conventional polynomially decreasing step size. Some important theoretical questions that deserve more attention are (i) why do stochastic momentum methods exhibit better generalization performance than SGD (Yan et al., 2018); (ii) how does the adaptive learning rate affect the generalization performance; (iii) how can we derive much sharper analysis of practical SGD for finding a global minimizer of deep learning with good generalization performance.

- **Better stochastic algorithms for deep learning.** Beyond theoretical questions mentioned above, it is also important to design better stochastic algorithms for deep learning. While most recent studies focus on designing better adaptive learning rates, however, they have mostly ignored the role of stochastic gradients itself. The learning rate plays its role through multiplying with stochastic gradients. We believe that it is important to consider the properties of stochastic gradients, which essentially depend on the data.

Non-Convex Min-Max Optimization and Generative Adversarial Networks. Recently, non-convex non-concave min-max optimization has received increasing attention due to its application in generative adversarial networks (GAN) (Goodfellow et al., 2014; Radford, Metz, & Chintala, 2015; Arjovsky, Chintala, & Bottou, 2017). GAN has emerged to be an important paradigm of unsupervised learning. It learns a generator network and a discriminator network in a unified framework by solving a min-max problem of the following form:

$$\min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{u}\in\mathcal{U}}\mathcal{L}(\mathbf{w},\mathbf{u}),$$

where $\mathbf{w}$ denotes the parameter of the generator network and $\mathbf{u}$ denotes the parameter of the discriminator network. Although many variants of GAN have been investigated, the research on optimization algorithms for GAN remains rare. In practice, most studies use a primal-dual variant of Adam for optimization, which runs several steps of Adam for updating the discriminator network and then runs one step of Adam for updating the generator network. Theoretically, most existing results of min-max optimization algorithms for GAN are either asymptotic (Daskalakis, Ilyas, Syrgkanis, & Zeng, 2017; Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017; Nagarajan & Kolter, 2017; Cherukuri, Gharesifard, & Cortes, 2017) or their analysis require strong assumptions of the problem (Nagarajan & Kolter, 2017; Grnarova, Levy, Lucchi, Hofmann, & Krause, 2017) (e.g., the problem is concave in maximization). In our recent study (Lin, Liu, Rafique, & Yang, 2018), we proposed new stochastic algorithms based on the proximal point framework for solving the non-convex non-concave min-max problem of GAN, and established their complexities for finding approximate first-order stationary points without convex and concavity assumptions.

Future studies in this direction should answer the following questions (i) how can we analyze the generalization performance of stochastic min-max optimization algorithms for GAN? (ii) does GAN exhibit some nice properties as in deep learning that facilitates the design of better stochastic algorithms? (iii) why is the Adam algorithm more effective than SGD for GAN? (iv) how can we design faster stochastic algorithms for solving non-convex non-concave min-max problems with lower complexities?

Other Non-Convex Learning Problems. Beyond regular deep learning and GAN, non-convex learning also has some important applications in machine learning. Below, we will mention several of them.

- **Learning with Non-convex Regularizers.** Learning with a non-convex regularizer can be formulated as:

  $$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) := \mathrm{E}_{\mathbf{z}}[f(\mathbf{w};\mathbf{z})] + R(\mathbf{w})$$

  where $R(\mathbf{w})$ denotes a regularizer, which includes the indicator function of a non-convex set. Commonly used non-convex regularizers that have been well studied include log-sum penalty (LSP) (Candès, Wakin, & Boyd, 2008), minimax concave penalty (MCP) (C.-H. Zhang, 2010), smoothly clipped absolute deviation (SCAD) (J. Fan & Li, 2001), capped ℓ_1 penalty (T. Zhang, 2010), transformed ℓ_1 norm (S. Zhang & Xin, 2014). However, there are many other interesting non-convex regularizations (Chartrand, 2012; Chartrand & Yin, 2016; Wen, Chu, Liu, & Qiu, 2018). For example, one can formulate learning a quantized neural network as a non-convex minimization with a non-convex constraint. Although non-smooth non-convex regularization has been considered in literature (Attouch, Bolte, & Svaiter, 2013; Bolte, Sabach, & Teboulle, 2014; Bot, Csetnek, & László, 2016; H. Li & Lin, 2015; Yu, Zheng, Marchetti-Bowick, & Xing, 2015; L. Yang, 2018; T. Liu, Pong, & Takeda, 2018; An & Nam, 2017; Zhong & Kwok, 2014), existing results are restricted to deterministic optimization and asymptotic or local convergence analysis. In our recent works (Y. Xu, Jin, & Yang, 2019; Y. Xu, Qi, Lin, Jin, & Yang, 2019), we have proposed new stochastic algorithms for tackling learning with a non-smooth non-convex regularizer, and established state-of-the-art non-asymptotic convergence rates.

- **DC programming.** Difference-of-Convex (DC) programming is to solve non-convex minimization problems of the following form:

  $$\min_{\mathbf{w}} f(\mathbf{w}) - g(\mathbf{w})$$

  where both f and g are convex functions. DC programming finds applications in many machine learning problems (Le Thi, Dinh, & Belghiti, 2014; Le Thi & Dinh, 2014; Nitanda & Suzuki, 2017; Thi, Le, Phan, & Tran, 2017; Khalaf, Astorino, d'Alessandro, & Gaudioso, 2017). For example, positive unlabeled learning problems can be formulated as a DC programming (Kiryo, Niu, du Plessis, & Sugiyama, 2017). In (Y. Xu, Qi, et al., 2019), we developed new stochastic DC algorithms for a broad family of DC problems, and established their complexities.

- **Distributionally Robust Optimization (DRO).** DRO is to solve the following min-max problem:

  $$\min_{\mathbf{w}\in\mathbb{R}^d} \max_{\mathbf{p}\in\mathcal{P}} \sum_{i=1}^{n} p_i f(\mathbf{w}, \mathbf{z}_i)$$

  where $\mathcal{P} \subseteq \{\mathbf{p} \in \mathbb{R}^n, \sum_i^n p_i = 1, p_i \geq 0\}$ encodes some constraint that how far $\mathbf{p}$ deviates from the empirical distribution $\hat{p}_i = 1/n, i = 1, \ldots, n$. DRO has found to be effective in handling imbalanced data (Namkoong & Duchi, 2016, 2017; Zhu, Li, Wang, Gong, & Yang, 2019; Y. Fan, Lyu, Ying, & Hu, 2017). It is also related to variance-based regularization and can yield smaller excess risk bounds (Namkoong & Duchi, 2017). When the loss function $f(\mathbf{w}, \mathbf{z})$ is non-convex in terms of $\mathbf{w}$, the above problem is non-convex and concave min-max problems. In (Rafique, Liu, Lin, & Yang, 2018), we have proposed efficient stochastic algorithms for solving the above min-max problems, and demonstrated that it gives better performance than SGD for learning a deep neural network in the presence of imbalanced data.

- **Learning with Truncated Losses.** Learning with truncated losses has long history in statistics (Wu & Liu, 2007; Belagiannis, Rupprecht, Carneiro, & Navab, 2015), which is more robust to outliers and can be formulated as

  $$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \phi(f(\mathbf{w}, \mathbf{z}_i))$$

  where $\phi(\cdot)$ is suitable concave truncation function (Y. Xu, Zhu, et al., 2019; Belagiannis et al., 2015). The above problem is a non-convex minimization problem. In (Y. Xu, Zhu, et al., 2019), we studied SGD for minimizing the above truncated losses and observed improved performance in the presence of various types of outliers and noise. However, it remains a question whether SGD converges to a global minimizer.

Heavily Constrained Learning

As ML is increasingly deployed in various domains, more and more problems are being formulated as constrained optimization problems where constraints are introduced to account for other factors/concerns beyond the prediction performance (B. Woodworth, Gunasekar, Ohannessian, & Srebro, 2017; Hardt et al., 2016; Globerson & Roweis, 2006b; Gupta et al., 2016; Canini et al., 2016; Globerson & Roweis, 2006b). Recently, there is much interest in measuring and ensuring fairness in ML, which is important in domains protected by anti-discrimination law (B. E. Woodworth et al., 2017; Hardt et al., 2016; Zafar et al., 2017; A. Agarwal et al., 2018). For example, a financial institution may want to use machine learning methods to predict whether a particular individual will pay back a loan or not for making a lending decision. In this case, it is morally and legally undesirable to discriminate based on the person's race and/or gender. A variety of notions of fairness has been considered in literature, including demographic parity, equality of opportunity, equalized odds, 80% rule, which can be modeled naturally as data dependent equality or inequality constraints (B. Woodworth et al., 2017; Hardt et al., 2016; Globerson & Roweis, 2006b).

Learning with data dependent constraints could also arise in *Interpretable learning*, which requires the prediction or the predictive model to be interpretable by a human. For example, if ML is used to predict whether a medication is effective for a client, then the client wants to know why it is effective in order to trust the medication. One way to achieve interpretable learning is to impose human-interpretable constraints into the learning process. For instance, for predicting an individual will pay back a loan or not, it is expected the probability of paying back is likely to increase as the person's income increases. It can be modeled as a constraint on the monotonicity of the predictive function respect to some features (Gupta et al., 2016).

Learning with complicated and complex constraints can find applications in other scenarios. In *Neyman-Pearson (NP) classification* paradigm (Rigollet & Tong, 2011), one needs to minimize false negative rate with an upper bound on false positive rate, where the upper bound on false positive rate is represented as a constraint. When the observed data are subject to some *uncertainty* (e.g, corruption, missing values, noise contamination), many studies have formulated the task as a constrained learning problem (Globerson & Roweis, 2006a; Sra et al., 2011). Recent works also found that imposing constrains on model parameters of neural networks can be more effective than using a regularization term in the objective for improving the prediction performance (Gouk, Frank, Pfahringer, & Cree, 2018; Ravi, Dinh, Lokhande, & Singh, 2018), and can improve the *robustness* of learned neural networks to adversarial examples (Cisse, Bojanowski, Grave, Dauphin, & Usunier, 2017). The robustness of a neural network is very important for applications in security critical domains (e.g., autonomous driving) (Carlini & Wagner, 2017; Tian, Pei, Jana, & Ray, 2018).

Constrained convex optimization has been studied extensively for a few decades and different methods, ranging from projected gradient methods, Frank-Wolfe methods (or conditional gradient methods), barrier methods, augmented Lagrangian methods, penalty methods, level-set methods to trust-region methods, have been developed and studied. However, the design of most existing constrained optimization algorithms suffers from severe scalability issues in the presence of big data and many complex constraints due to various reasons.

The general constrained learning problem can be formulated as:

$$\min_{\mathbf{x}\in\mathcal{X}} f_0(\mathbf{x}), \tag{2}$$

$$s.t.\ f_i(\mathbf{x}) \le r_i, i = 1, \ldots, m \tag{3}$$

In (Mahdavi, Yang, Jin, & Zhu, 2012; T. Yang, Lin, & Zhang, 2017), we developed new theories of projection reduced (stochastic) first-order methods with only one or a logarithmic number of projections. In (Lin, Nadarajah, Soheili, & Yang, 2019), we developed new stochastic level-set methods for a family of finite-sum constrained convex optimization problems which can guarantee the exact feasibility of constraints. Recently, we proposed a class of subgradient methods for constrained optimization where the objective function and the constraint functions are non-convex (Ma, Lin, & Yang, 2019).

However, there still remain many challenging problems for heavily constrained learning.

- How to efficiently handle a large number of constraints?
- How do the constraints affect the generalization performance of a learned model?
- How to establish stronger convergence for a constrained optimization with non-convex objectives and non-convex constraints?

Acknowledgments

We acknowledge the support by National Science Foundation Career Award 1844403. Part of the our research cited in this paper was supported by National Science Foundation (IIS-1545995).

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In *Proceedings of the 35th international conference on machine learning (icml)* (pp. –).

Agarwal, N., Allen Zhu, Z., Bullins, B., Hazan, E., & Ma, T. (2017). Finding approximate local minima faster than gradient descent. In *Acm symposium on theory of computing (stoc)* (pp. 1195–1199).

Allen-Zhu, Z., Li, Y., & Song, Z. (2018). A convergence theory for deep learning via over-parameterization. *CoRR*, *abs/1811.03962*.

Allen-Zhu, Z. (2017). Natasha 2: Faster non-convex optimization than sgd. *CoRR*, */abs/1708.08694/v4*.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., . . . Zhu, Z. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd international conference on international conference on machine learning (icml)* (pp. 173–182).

An, N. T., & Nam, N. M. (2017). Convergence analysis of a proximal point algorithm for minimizing differences of functions. *Optimization*, *66*(1), 129-147.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214–223).

Arora, S., Cohen, N., & Hazan, E. (2018). On the optimization of deep networks: Implicit acceleration by overparameterization. *arXiv preprint arXiv:1802.06509*.

Attouch, H., Bolte, J., & Svaiter, B. F. (2013). Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, *137*(1), 91–129.

Belagiannis, V., Rupprecht, C., Carneiro, G., & Navab, N. (2015). Robust optimization for deep regression. In *Proceedings of the ieee international conference on computer vision* (pp. 2830–2838).

Bolte, J., Sabach, S., & Teboulle, M. (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, *146*, 459–494.

Bot, R. I., Csetnek, E. R., & László, S. C. (2016, Feb 01). An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. *EURO Journal on Computational Optimization*, *4*(1), 3–25.

Candès, E. J., Wakin, M. B., & Boyd, S. P. (2008, Dec 01). Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, *14*(5), 877–905.

Canini, K., Cotter, A., Gupta, M. R., Fard, M. M., & Pfeifer, J. (2016). Fast and flexible monotonic functions with ensembles of lattices. In *Proceedings of the 30th international conference on neural information processing systems (nips)* (pp. 2927–2935).

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)* (pp. 39–57).

Carmon, Y., Duchi, J. C., Hinder, O., & Sidford, A. (2016). Accelerated methods for non-convex optimization. *CoRR*, *abs/1611.00756*.

Cartis, C., Gould, N. I. M., & Toint, P. L. (2011a, Dec 01). Adaptive cubic regularisation methods for unconstrained optimization. part ii: worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, *130*(2), 295–319.

Cartis, C., Gould, N. I. M., & Toint, P. L. (2011b). Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Mathematical Programming*, *127*(2), 245–295.

Chartrand, R. (2012). Nonconvex splitting for regularized low-rank+ sparse decomposition. *IEEE Transactions on Signal Processing*, *60*(11), 5810–5819.

Chartrand, R., & Yin, W. (2016). Nonconvex sparse regularization and splitting algorithms. In *Splitting methods in communication, imaging, science, and engineering* (pp. 237–249). Springer.

Chen, J., & Gu, Q. (2018). Closing the generalization gap of adaptive gradient methods in training deep neural networks.

arXiv preprint arXiv:1806.06763.

Chen, Z., Yuan, Z., Yi, J., Zhou, B., Chen, E., & Yang, T. (2019). Universal stagewise learning for non-convex problems with convergence on averaged solutions. In *7th international conference on learning representations, ICLR 2019, new orleans, la, usa, may 6-9, 2019.*

Cherukuri, A., Gharesifard, B., & Cortes, J. (2017). Saddle-point dynamics: conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, *55*(1), 486–511.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., & Usunier, N. (2017). Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 854–863).

Daskalakis, C., Ilyas, A., Syrgkanis, V., & Zeng, H. (2017). Training gans with optimism. *CoRR*, *abs/1711.00141*.

Davis, D., & Drusvyatskiy, D. (2018). Stochastic subgradient method converges at the rate o($k^{-1/4}$) on weakly convex functions. *arXiv preprint arXiv:1802.02988*.

Deng, L., & Liu, Y. (2018). *Deep learning in natural language processing*. Springer.

Du, S. S., Zhai, X., Poczos, B., & Singh, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121–2159.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, Y., Lyu, S., Ying, Y., & Hu, B. (2017). Learning with average top-k loss. In *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, 4-9 december 2017, long beach, ca, USA* (pp. 497–505).

Ghadimi, S., & Lan, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, *23*(4), 2341–2368.

Globerson, A., & Roweis, S. (2006a). Nightmare at test time: Robust learning by feature deletion. In *Proceedings of the 23rd international conference on machine learning* (pp. 353–360).

Globerson, A., & Roweis, S. (2006b). Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd international conference on machine learning* (pp. 353–360).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Gouk, H., Frank, E., Pfahringer, B., & Cree, M. (2018). Regularisation of neural networks by enforcing lipschitz continuity. *arXiv preprint arXiv:1804.04368*.

Grnarova, P., Levy, K. Y., Lucchi, A., Hofmann, T., & Krause, A. (2017). An online learning approach to generative adversarial networks. *CoRR*, *abs/1706.03269*.

Gupta, M. R., Cotter, A., Pfeifer, J., Voevodski, K., Canini, K. R., Mangylov, A., ... Esbroeck, A. V. (2016). Monotonic calibrated interpolated look-up tables. *Journal of Machine Learning Research (JMLR)*, *17*, 109:1–109:47.

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315–3323).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems 30 (nips)* (pp. 6629–6640).

Hillar, C. J., & Lim, L.-H. (2013, November). Most tensor problems are np-hard. *Journal of ACM*, *60*(6), 45:1–45:39.

Khalaf, W., Astorino, A., d'Alessandro, P., & Gaudioso, M. (2017). A dc optimization-based clustering technique for edge detection. *Optimization Letters*, *11*(3), 627–640.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings.* Retrieved from http://arxiv.org/abs/1412.6980

Kiryo, R., Niu, G., du Plessis, M. C., & Sugiyama, M. (2017). Positive-unlabeled learning with non-negative risk estimator. In *Advances in neural information processing systems 30* (pp. 1675–1685).

Kohler, J. M., & Lucchi, A. (2017). Sub-sampled cubic regularization for non-convex optimization. In *Proceedings of the international conference on machine learning (icml)* (pp. 1895–1904).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (nips)* (pp. 1106–1114).

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Le Thi, H. A., & Dinh, T. P. (2014). Dc programming in communication systems: challenging problems and methods. *Vietnam Journal of Computer Science*, *1*(1), 15–28.

Le Thi, H. A., Dinh, T. P., & Belghiti, M. (2014). Dca based algorithms for multiple sequence alignment (msa). *Central European Journal of Operations Research*, *22*(3), 501–524.

Li, H., & Lin, Z. (2015). Accelerated proximal gradient methods for nonconvex programming. In *Proceedings of the 28th international conference on neural information processing systems - volume 1* (pp. 379–387).

Li, X., & Orabona, F. (2018). On the convergence of stochastic gradient descent with adaptive stepsizes. *arXiv preprint arXiv:1805.08114*.

Li, Y., & Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in neural information processing systems (neurips)* (pp. 8157–8166).

Lin, Q., Liu, M., Rafique, H., & Yang, T. (2018). Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207*.

Lin, Q., Nadarajah, S., Soheili, N., & Yang, T. (2019). A data efficient and feasible level set method for stochastic convex optimization with expectation constraints. *CoRR*, *abs/1908.03077*.

Liu, M., & Yang, T. (2017a). On noisy negative curvature descent: Competing with gradient descent for faster non-convex optimization. *CoRR*, *abs/1709.08571*.

Liu, M., & Yang, T. (2017b). Stochastic non-convex optimization with strong high probability second-order convergence. *CoRR*, *abs/1710.09447*.

Liu, T., Pong, T. K., & Takeda, A. (2018, Sep 08). A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems. *Mathematical Programming*.

Luo, L., Xiong, Y., Liu, Y., & Sun, X. (2019). Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*.

Ma, R., Lin, Q., & Yang, T. (2019). Proximally constrained methods for weakly convex optimization with weakly convex constraints. *arXiv preprint arXiv:1908.01871*.

Mahdavi, M., Yang, T., Jin, R., & Zhu, S. (2012). Stochastic gradient descent with only one projection. In *Advances in neural information processing systems (nips)* (p. 503-511).

Nagarajan, V., & Kolter, J. Z. (2017). Gradient descent GAN optimization is locally stable. In *Advances in neural information processing systems 30 (nips)* (pp. 5591–5600).

Namkoong, H., & Duchi, J. C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in neural information processing systems* (pp. 2208–2216).

Namkoong, H., & Duchi, J. C. (2017). Variance-based regularization with convex objectives. In *Advances in neural information processing systems* (pp. 2971–2980).

Nesterov, Y., & Polyak, B. T. (2006). Cubic regularization of newton method and its global performance. *Math. Program.*, *108*(1), 177–205.

Nitanda, A., & Suzuki, T. (2017). Stochastic difference of convex algorithm and its application to training deep boltzmann machines. In *Artificial intelligence and statistics* (pp. 470–478).

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rafique, H., Liu, M., Lin, Q., & Yang, T. (2018). Non-convex min-max optimization: Provable algorithms and applications in machine learning. *CoRR*, *abs/1810.02060*.

Ravi, S. N., Dinh, T., Lokhande, V. S. R., & Singh, V. (2018). Constrained deep learning using conditional gradient and applications in computer vision. *arXiv preprint arXiv:1803.06453*.

Real, E., Aggarwal, A., Huang, Y., & Le, Q. V. (2019). Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 4780–4789).

Reddi, S. J., Zaheer, M., Sra, S., Poczos, B., Bach, F., Salakhutdinov, R., & Smola, A. J. (2017). A generic approach for escaping saddle points. *arXiv preprint arXiv:1709.01434*.

Rigollet, P., & Tong, X. (2011, November). Neyman-pearson classification, convexity and stochastic constraints. *J. Mach. Learn. Res.*, *12*, 2831–2855.

Royer, C. W., & Wright, S. J. (2017). Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *CoRR*, *abs/1706.03131*.

Sra, S., Nowozin, S., & Wright, S. J. (2011). *Optimization for machine learning*. The MIT Press.

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

Thi, H. A. L., Le, H. M., Phan, D. N., & Tran, B. (2017). Stochastic dca for the large-sum of non-convex functions problem and its application to group variable selection in classification. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 3394–3403).

Tian, Y., Pei, K., Jana, S., & Ray, B. (2018). Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th international conference on software engineering* (pp. 303–314).

Wen, F., Chu, L., Liu, P., & Qiu, R. C. (2018). A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. *IEEE Access*, *6*, 69883–69906.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., & Srebro, N. (2017). Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*.

Woodworth, B. E., Gunasekar, S., Ohannessian, M. I., & Srebro, N. (2017). Learning non-discriminatory predictors. In *Proceedings of the 30th conference on learning theory, COLT 2017, amsterdam, the netherlands, 7-10 july 2017* (pp. 1920–1953).

Wu, Y., & Liu, Y. (2007). Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, *102*(479), 974–983.

Xu, P., Roosta-Khorasani, F., & Mahoney, M. W. (2017). Newton-type methods for non-convex optimization under inexact hessian information. *CoRR*, *abs/1708.07164*.

Xu, Y., Jin, R., & Yang, T. (2019). Stochastic proximal gradient methods for non-smooth non-convex regularized problems. *arXiv preprint arXiv:1902.07672*.

Xu, Y., Lin, Q., & Yang, T. (2017). Stochastic convex optimization: Faster local growth implies faster global convergence. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 3821–3830).

Xu, Y., Qi, Q., Lin, Q., Jin, R., & Yang, T. (2019). Stochastic optimization for DC functions and non-smooth non-convex regularizers with non-asymptotic convergence. In *Proceedings of the 36th international conference on machine learning, ICML 2019, 9-15 june 2019, long beach, california, USA* (pp. 6942–6951).

Xu, Y., Rong, J., & Yang, T. (2018). First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Advances in neural information processing systems (neurips)* (pp. 5530–5540).

Xu, Y., Zhu, S., Yang, S., Zhang, C., Jin, R.,

& Yang, T. (2019). Learning with nonconvex truncated losses by SGD. In *Proceedings of the thirty-fifth conference on uncertainty in artificial intelligence, UAI 2019, tel aviv, israel, july 22-25, 2019* (p. 244).

Yan, Y., Yang, T., Li, Z., Lin, Q., & Yang, Y. (2018). A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI 2018, july 13-19, 2018, stockholm, sweden.* (pp. 2955–2961).

Yang, L. (2018). Proximal gradient method with extrapolation and line search for a class of nonconvex and nonsmooth problems. *CoRR*, *abs/1711.06831*.

Yang, T., Lin, Q., & Zhang, L. (2017). A richer theory of convex constrained optimization with reduced projections and improved rates. In *Proceedings of the 34th international conference on machine learning (icml)* (p. -).

Yang, T., Mahdavi, M., Jin, R., Zhang, L., & Zhou, Y. (2012). Multiple kernel learning from noisy labels by stochastic programming. In *Proceedings of the international conference on machine learning (icml)* (pp. 233–240).

You, S., Ding, D., Canini, K. R., Pfeifer, J., & Gupta, M. R. (2017). Deep lattice networks and partial monotonic functions. In *Advances in neural information processing systems 30 (nips)* (pp. 2985–2993).

Yu, Y., Zheng, X., Marchetti-Bowick, M., & Xing, E. P. (2015). Minimizing nonconvex non-separable functions. In *The 17^{th} international conference on artificial intelligence and statistics (AISTATS).*

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web* (pp. 1171–1180).

Zaheer, M., Reddi, S., Sachan, D., Kale, S., & Kumar, S. (2018). Adaptive methods for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31* (pp. 9793–9803). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/8186-adaptive-methods-for-nonconvex-optimization.pdf

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894 – 942.

Zhang, C.-H., & Zhang, T. (2012, 11). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, *27*(4), 576–593. doi: 10.1214/12-STS399

Zhang, S., & Xin, J. (2014). Minimization of transformed l_1 penalty: Theory, difference of convex function algorithm, and robust application in compressed sensing. *CoRR*, *abs/1411.5735*.

Zhang, T. (2010, March). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.*, *11*, 1081–1107.

Zhong, W., & Kwok, J. T. (2014). Gradient descent with proximal average for nonconvex and composite regularization. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence, july 27 -31, 2014, québec city, québec, canada.* (pp. 2206–2212).

Zhou, D., Tang, Y., Yang, Z., Cao, Y., & Gu, Q. (2018). On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*.

Zhu, D., Li, Z., Wang, X., Gong, B., & Yang, T. (2019). A robust zero-sum game framework for pool-based active learning. In *The 22nd international conference on artificial intelligence and statistics* (pp. 517–526).

Zou, D., Cao, Y., Zhou, D., & Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep relu networks. *CoRR*, *abs/1811.08888*.

Tianbao Yang is an assistant professor at the Computer Science Department at the University of Iowa since 2014. He was a researcher at NEC Laboratories America, and a researcher at GE Global Research before joining UIowa. He has won the best student paper award at COLT 2012 and the NSF Career Award. He is an associate editor for Neurocomputing Journal, and the Journal of Mathematical Foundations of Computing. He has served as senior program committee member for AAAI and IJCAI and reviewer for AISTATS, ICML and NeurIPS.

Considerations for AI Fairness for People with Disabilities

Shari Trewin (IBM; trewin@us.ibm.com)
Sara Basson (Google Inc.; basson@google.com)
Michael Muller (IBM Research; michael_muller@us.ibm.com)
Stacy Branham (University of California, Irvine; sbranham@uci.edu)
Jutta Treviranus (OCAD University; jtreviranus@ocadu.ca)
Daniel Gruen (IBM Research; daniel_gruen@us.ibm.com)
Daniel Hebert (IBM; dhebert@us.ibm.com)
Natalia Lyckowski (IBM; nladuca@us.ibm.com)
Erich Manser (IBM; emanser@us.ibm.com)

DOI: 10.1145/3362077.3362086

Abstract

In society today, people experiencing disability can face discrimination. As artificial intelligence solutions take on increasingly important roles in decision-making and interaction, they have the potential to impact fair treatment of people with disabilities in society both positively and negatively. We describe some of the opportunities and risks across four emerging AI application areas: employment, education, public safety, and healthcare, identified in a workshop with participants experiencing a range of disabilities. In many existing situations, non-AI solutions are already discriminatory, and introducing AI runs the risk of simply perpetuating and replicating these flaws. We next discuss strategies for supporting fairness in the context of disability throughout the AI development lifecycle. AI systems should be reviewed for potential impact on the user in their broader context of use. They should offer opportunities to redress errors, and for users and those impacted to raise fairness concerns. People with disabilities should be included when sourcing data to build models, and in testing, to create a more inclusive and robust system. Finally, we offer pointers into an established body of literature on human-centered design processes and philosophies that may assist AI and ML engineers in innovating algorithms that reduce harm and ultimately enhance the lives of people with disabilities.

Introduction

Systems that leverage Artificial Intelligence are becoming pervasive across industry sectors (Costello, 2019), as are concerns that these technologies can unintentionally exclude or lead to unfair outcomes for marginalized populations (Bird, Hutchinson, Kenthapadi, Kiciman, & Mitchell, 2019)(Cutler, Pribik, & Humphrey, 2019)(IEEE & Systems, 2019)(Kroll et al., 2016)(Lepri, Oliver, Letouzé, Pentland, & Vinck, 2018). Initiatives to improve AI fairness for people across racial (Hankerson et al., 2016), gender (Hamidi, Scheuerman, & Branham, 2018a), and other identities are emerging, but there has been relatively little work focusing on AI fairness for people with disabilities. There are numerous examples of AI that can empower people with disabilities, such as autonomous vehicles (Brewer & Kameswaran, 2018) and voice agents (Pradhan, Mehta, & Findlater, 2018) for people with mobility and vision impairments. However, AI solutions may also result in unfair outcomes, as when Idahoans with cognitive/learning disabilities had their healthcare benefits reduced based on biased AI (*K.W. v. Armstrong, No. 14-35296 (9th Cir. 2015) :: Justia*, 2015). These scenarios suggest that the prospects of AI for people with disabilities are promising yet fraught with challenges that require the sort of upfront attention to ethics in the development process advocated by scholars (Bird et al., 2019) and practitioners (Cutler et al., 2019).

The challenges of ensuring AI fairness in the context of disability emerge from multiple sources. From the very beginning of algorithmic development, in the problem scoping stage, bias can be introduced by lack of awareness of the experiences and use cases of people with disabilities. Since systems are predicated on data, in data sourcing and data pre-processing stages, it is critical to gather data that include people with disabilities and to ensure that these data are not completely subsumed by data from presumed "normative" populations. This leads to a potential conundrum. The data need to be gathered in order to be reflected in the models, but confidentiality and privacy, especially as regards disability status, might make collecting these data difficult (for developers) or dangerous (for subjects) (Faucett, Ringland, Cullen, & Hayes, 2017)(von Schrader, Malzer, & Bruyère, 2014). Another area to address during model training and testing is the potential for model bias. Owing to intended or unintended bias in the data, the model may inadvertently enforce or reinforce discriminatory patterns that work against people with disabilities (Janssen & Kuk, 2016). We advocate for increased awareness of these patterns, so we can avoid replication of past bias into future algorithmic decisions, as has been well-documented in banking (Bruckner, 2018)(Chander, 2017)(Hurley & Adebayo, 2016). Finally, once a trained model is incorporated in an application, it is then critical to test with diverse users, particularly those deemed as outliers. This paper provides a number of recommendations towards overcoming these challenges.

In the remainder of this article, we overview the nascent area of AI Fairness for People with Disabilities as a practical pursuit and an academic discipline. We provide a series of examples that demonstrate the potential for harm to people with disabilities across four emerging AI application areas: employment, education, public safety, and healthcare. Then, we identify strategies of developing AI algorithms that resist reifying systematic societal exclusions at each stage of AI development. Finally, we offer pointers into an established body of literature on human-centered design processes and philosophies that may assist AI and ML engineers in innovating algorithms that reduce harm and – as should be our ideal – ultimately enhance the lives of people with disabilities.

Related Work

The 2019 Gartner CIO survey (Costello, 2019) of 3000 enterprises across major industries reported that 37% have implemented some form of AI solution, an increase of 270% over the last four years. In parallel, there is increasing recognition that intelligent systems should be developed with attention to the ethical aspects of their behavior (Cutler et al., 2019)(IEEE & Systems, 2019), and that fairness should be considered upfront, rather than as an afterthought (Bird et al., 2019). IEEE's Global Initiative on Ethics of Autonomous and Intelligent Systems is developing a series of international standards for such processes (Koene, Smith, Egawa, Mandalh, & Hatada, 2018), including a process for addressing ethical concerns during design (P7000), and the P7003 Standard for Algorithmic Bias Considerations (Koene, Dowthwaite, & Seth, 2018). There is ongoing concern and discussion about accountability for potentially harmful decisions made by algorithms(Kroll et al., 2016)(Lepri et al., 2018), with some new academic initiatives – like one at Georgetown's Institute for Tech Law & Policy (Givens, 2019), and a workshop at the ASSETS 2019 conference(Trewin et al., 2019) – focusing specifically on AI and Fairness for People with Disabilities.

Any algorithmic decision-process can be biased, and the FATE/ML community is actively developing approaches for detection and remediation of bias (Kanellopoulos, 2018)(Lohia et al., 2019). Williams, Brooks and Shmargad show how racial discrimination can arise in employment and education even without having social category information, and how the lack of category information makes such biases harder to detect (Williams, Brooks, & Shmargad, 2018). Although they argue for inclusion of social category information in algorithmic decision-making, they also acknowledge the potential harm that can be caused to an individual by revealing sensitive social data such as immigration status. Selbst et al. argue that purely algorithmic approaches are not sufficient, and the full social context of deployment must be considered if fair outcomes are to be achieved (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019).

Some concerns about AI fairness in the

context of individuals with disabilities or neurological or sensory differences are now being raised (Fruchterman & Mellea, 2018)(Guo, Kamar, Vaughan, Wallach, & Morris, 2019)(Lewis, 2019)(Treviranus, 2019)(Trewin, 2018a), but research in this area is sparse. Fruchterman and Mellea (Fruchterman & Mellea, 2018) outline the widespread use of AI tools in employment and recruiting, and highlight some potentially serious implications for people with disabilities, including the analysis of facial movements and voice in recruitment, personality tests that disproportionately screen out people with disabilities, and the use of variables that could be discriminatory, such as gaps in employment. "Advocates for people with disabilities should be looking at the proxies and the models used by AI vendors for these "hidden" tools of discrimination" (Fruchterman & Mellea, 2018).

Motivating Examples

In October 2018, a group of 40 disability advocates, individuals with disabilities, AI and accessibility researchers and practitioners from industry and academia convened in a workshop (Trewin, 2018b) to identify and discuss the topic of fairness for people with disabilities in light of the increasing mainstream application of AI solutions in many industries (Costello, 2019). This section describes some of the opportunities and risks identified by the workshop participants in the areas of employment, education, public safety and healthcare.

Employment

People with disabilities are no strangers to discrimination in hiring practices. In one recent field study, disclosing a disability (spinal cord injury or Asperger's Syndrome) in a job application cover letter resulted in 26% fewer positive responses from employers, even though the disability was not likely to affect productivity for the position (Ameri et al., 2018). When it comes to inclusive hiring, it has been shown that men and those who lack experience with disability tend to have more negative affective reactions to working with individuals with disabilities (Popovich, Scherbaum, Scherbaum, & Polinko, 2003). Exclusion can be unintentional. For example, qualified deaf candidates who speak through an interpreter may be screened out for a position requiring verbal communication skills, even though they could use accommodations to do the job effectively. Additional discriminatory practices are particularly damaging to this population, where employment levels are already low: In 2018, the employment rate for people with disabilities was 19.1%, while the employment percentage for people without disabilities was 65.9% (Bureau of Labor Statistics, 2019).

Employers are increasingly relying on technology in their hiring practices. One of their selling points is the potential to provide a fairer recruitment process, not influenced by an individual recruiter's bias or lack of knowledge. Machine learning models are being used for candidate screening and matching jobseekers with available positions. There are AI-driven recruitment solutions on the market today that analyze online profiles and resumes, the results of online tests, and video interview data, all of which raise potential concerns for disability discrimination (Fruchterman & Mellea, 2018). While the use of AI in HR and recruitment is an increasing trend (Faggella, 2019), there are already cautionary incidents of discrimination, as when Amazon's AI recruiting solution "learned" to devalue resumes of women (Dastin, 2018).

The workshop identified several risk scenarios:

- A deaf person may be the first person using sign language interpretation to apply to an organization. Candidate screening models that learn from the current workforce will perpetuate the status quo, and the biases of the past. They will likely exclude candidates with workplace differences, including those who use accommodations to perform their work.
- An applicant taking an online test using assistive technology may take longer to answer questions, especially if the test itself has not been well designed for accessibility. Models that use timing information may end up systematically excluding assistive technology users. Resumes and job applications may not contain explicit information about a person's disability, but other variables may be impacted, including gaps

in employment, school attended, or time to perform an online task.

- An applicant with low facial affect could be screened out by a selection process that uses video analysis of eye gaze, voice characteristics, or facial movements, even though she is highly skilled. This type of screening poses a barrier for anyone whose appearance, voice or facial expression differs from the average. It could exclude autistic individuals or blind applicants who do not make eye contact, deaf people and others who do not communicate verbally, people with speech disorders or facial paralysis, or people who have survived a stroke, to name a few.

When the available data do not include many people with disabilities, and reflect existing biases, and the deployed systems rely on proxies that are impacted by disability, the risk of unfair treatment in employment is significant. We must seek approaches that do not perpetuate the biases of the past, or introduce new barriers by failing to recognize qualified candidates because they are different, or use accommodations to do their work.

Education

In the United States, people with disabilities have historically been denied access to free public education (Dudley-Marling & Burns, 2014)(Obiakor, Harris, Mutua, Rotatori, & Algozzine, 2012). It was nearly 20 years after the passing of Brown v. Board of Education, which desegregated public schools along racial lines, that the Education for All Handicapped Children Act was passed (Dudley-Marling & Burns, 2014), mandating that all students are entitled to a “free and appropriate public education” in the “least restrictive environment.” Prior to 1975, a mere one in five learners with disabilities had access to public school environments, often in segregated classrooms (Dudley-Marling & Burns, 2014). Despite great strides, some learners with disabilities still cannot access integrated public learning environments (Dudley-Marling & Burns, 2014), K-12 classroom technologies are often inaccessible (Shaheen & Lazar, 2018), postsecondary online learning materials are often inaccessible (Burgstahler, 2015) (Straumsheim, 2017), and e-learning platforms do not consider the needs of all learners (Cinquin, Guitton, & Sauzéon, 2019).

AI in the education market is being driven by the rapid transition from onsite classroom based education to online learning. Institutions can now expand their online learning initiatives to reach more students in a cost-effective manner. Industry analyst, Global Market Insights (Bhutani & Wadhwani, 2019), predicts that the market will grow to a $6 Billion dollar industry by 2024. The new generation of online learning platforms are integrated with AI technologies and use them to personalize learning (and testing) for each student, among other applications. Two examples of providers of these systems are from traditional Learning Management System (LMS) vendors like Blackboard; and more recently from the Massive Open Online Course (MooC) providers like edX.

Personalized learning could provide enormous benefits for learners with disabilities, e.g. (Morris, Kirschbaum, & Picard, 2010). It could be as simple as augmenting existing content with additional illustrations and pictures for students who are classified as visual learners, or as complex as generation of personalized user interfaces (Gajos, Wobbrock, & Weld, 2007). For non-native language speakers, including deaf learners, the system could provide captions for video content so the student can read along with the lecture.

Any system that makes inferences about a student’s knowledge and abilities based on their online interactions runs the risk of misinterpreting and underestimating students with disabilities. Students whose learning capabilities or styles are outside the presumed norm may not receive fair treatment. For example, if there is a rigid time constraint for completing a test or a quiz, a student that has a cognitive disability or test anxiety where they process information more slowly than other students would be assessed as being less capable than they are.

Unlike other areas, in an educational setting, disability information may often be available, and the challenge is to provide differentiated education for a wide range of people, without introducing bias against disability groups.

Public Safety

People with disabilities are much more likely to be victims of violent crime than people without disabilities (Harrell, 2017). Threats come not only from other citizens, but also from law enforcement itself; for example, police officers can misinterpret people with disabilities as being uncooperative or even threatening, and deprive them of access to Miranda warnings (US Department of Justice Civil Rights Division, 2006). Law enforcement's implicit bias and discrimination towards people with disabilities, as well as the potential for technology to address these challenges, are both featured in the 2015 Final Report of the President's Task Force on 21st Century Policing (Policing, 2015).

The application of AI technology to identify threats to public safety and enforce the law is highly controversial (McCarthy, 2019). This includes technology for identifying people, recognizing individuals, and for interpreting behavior (for example, whether someone is acting in a suspicious manner). Aside from the threat to personal privacy, the potential for errors and biased performance is very real. While public discourse and academic attention has so far focused on racial and gender disparities, workshop participants identified serious concerns and also some opportunities for people with disabilities.

Autonomous vehicles must be able to identify people in the environment with great precision. They must reliably recognize individuals who use different kinds of wheelchair and mobility devices, or move in an unusual way. One workshop participant described a wheelchair-using friend who propels themselves backwards with their feet. This is an unusual method of getting around, but the ability to recognize and correctly identify such outlier individuals is a matter of life and death.

Another participant had recently observed, a dishevelled man pacing restlessly in an airport lounge, muttering to himself, clearly in a state of high stress. His behavior could be interpreted by both humans and AI analysis as a potential threat. Instead he may be showing signs of an anxiety disorder, autism, or a strong fear of flying. Deaf signers' strong facial expressions can be misinterpreted (Shaffer & Rogan, 2018), leading to them being wrongly identified as being angry, and a potential security threat. Someone with an altered gait could be using a prosthesis, not hiding a weapon.

People with cognitive disabilities may be at especially high risk of being misidentified as a potential threat. Combining this with the need to respond quickly to genuine threats creates a dangerous situation and requires careful design of the AI system and its method of deployment.

There may also be opportunities for AI to improve public safety for people with disabilities. For example, AI-based interpretation could be trained to 'understand' a wide range of behaviors including hand flapping, pacing and sign language, as normal. A recent survey and interview study of people who are blind (Branham et al., 2017) suggests that facial and image recognition technologies could better support personal safety for individuals with sensory disabilities. They may support locating police officers and identifying fraudulent actors claiming to be officials. They may allow a person who is blind or deaf to be made aware of a weapon being brandished or discharged. They may support access to facial cues for more cautious and effective interactions with a potential aggressor or a police officer. These technologies may even allow blind individuals to provide more persuasive evidence to catch their perpetrators.

When considering the ethics of proposed projects in this space, the potential risks for individuals with disabilities should also be evaluated and addressed in the overall design. For example, a system could highlight someone with an altered gait, and list possibilities as someone hiding a weapon, or someone using a prosthesis or mobility device. In a situation where facial recognition is being used, a person's profile could help responders to avoid misunderstanding, but again this comes at the cost of sacrificing privacy, and potentially doing harm to other marginalized groups (Hamidi et al., 2018a). An overall balance must be found between using AI as a tool for maintaining public safety while minimizing negative outcomes for vulnerable groups and outlier individuals.

Healthcare

Today there are large disparities in access to healthcare for people with disabilities (Iezzoni, 2011) (Krahn, Walker, & Correa-De-Araujo, 2015), especially those with developmental disabilities (Krahn, Hammond, & Turner, 2006). Patients who are non-verbal or patients with cognitive impairments are often under-served or incorrectly served (Barnett, McKee, Smith, & Pearson, 2011) (Krahn & Fox, 2014) (Krahn et al., 2015). Deaf patients are often misdiagnosed with having a mental illness or a disorder, because of lack of cultural awareness or language barrier (Glickman, 2007) (Pollard, 1994). Another area that is underserved in the current model are patients with rare diseases or genetic disorders, that do not fall within standard protocols (Wastfelt, Fadeel, & Henter, 2006). For older adults with deteriorating health, this may lead to unwanted institutionalization. Promising technological developments, many of which include AI, abound, but need to better incorporate target users in the development process (Haigh & Yanco, 2002).

AI applications in healthcare could help to overcome some of the barriers preventing people getting access to the care and preventative care they need. For example, a non-verbal person may have difficulty communicating a problem they are experiencing. With respect to pain management or prescription delivery, AI can remove the requirement that patients advocate on their own behalf. For complex cases where disabilities or communicative abilities may affect treatment and ability to adhere to a treatment plan, AI could be applied to recognize special needs situations and flag them for extra attention, and build a case for a suitable course of treatment. With respect to rare diseases or genetic disorders, disparate data points can be aggregated such that solution determination and delivery is not contingent on an individual practitioner's know-how.

Unfortunately, there are no standards or regulations to assess the safety and efficacy of these systems. If the datasets don't well-represent the broader population, AI might work less well where data are scarce or difficult to collect. This can negatively impact people with rare medical conditions/disabilities. For example, if speech pauses are used to diagnose conditions like Alzheimer's disease, a person whose speech is already affected by a disability may be wrongly diagnosed, or their diagnosis may be missed because the system does not work for them. Pauses in speech can be because person is a non-native speaker; and not a marker of disease.

Just as we have seen in the domains of employment, education, and public safety, if healthcare applications are built for the extremes of currently excluded populations, the solution stands to improve fairness in access, instead of locking people out. Across all domains, AI applications pose both risks and opportunities for people with disabilities. The question remains: how and when can fairness for people with disabilities be implemented in the software development process towards minimizing risks and maximizing benefits? In the following section, we address this question for each stage of the AI development process.

Considerations for AI Practitioners

In this section, we recommend ways AI practitioners can be aware of, and work towards fairness and inclusion for people with disabilities in their AI-based applications. The section is organized around the typical stages of AI model development: problem scoping, data sourcing, pre-processing, model selection and training, and incorporating AI in an application.

Problem Scoping

Some projects have greater potential to impact human lives than others. To identify areas where special attention may need to be paid to fairness, it can be helpful to apply the Bioss AI Protocol (Bioss, 2019), which recommends asking the following 5 questions about the work being done by AI:

1. Is the work **Advisory**, leaving space for human judgement and decision making?
2. Has the AI been granted any **Authority** over people?
3. Does the AI have **Agency** (the ability to act in a given environment)?
4. What skills and responsibilities are we at risk of **Abdicating**?

5. Are lines of **Accountability** clear, in what are still organizations run by human beings?

AI practitioners can also investigate whether this is an area where people with disabilities have historically experienced discrimination, such as employment, housing, education, and healthcare. If so, can the project improve on the past? Identify what specific outcomes there should be, so these can be checked as the project progresses. Develop a plan for tackling bias in source data to avoid perpetuating previous discriminatory treatment. This could include boosting representation of people with disabilities, adjusting for bias against specific disability groups, or flagging gaps in data coverage so the limits of the resulting model are explicit.

Approaches to developing ethical AI include actively seeking the ongoing involvement of a diverse set of stakeholders (Cutler et al., 2019), and a diversity of data to work with. To extend this approach to people with disabilities, it may be useful to define a set of 'outlier' individuals, and include them in the team, following an inclusive design method as discussed in the following section. These are people whose data may look very different to the average. What defines an outlier depends on the application. Many variables can be impacted by a disability, leading to a potential for bias even where no explicit disability information is available. For example, in speech recognition it could be a person with a stutter or a person with slow, slurred speech. In a healthcare application involving height, this could mean including a person of short stature. Outliers may also include people who belong with one group, but whose data looks more like that of another group. For example, a person who is slow to take a test may not be struggling with the material, but with typing, or accessing the test itself through their assistive technology. By defining outlier individuals up front, the design process can consider at each stage what their needs are, whether there are potential harms that need to be avoided, and how to achieve this.

Related to identifying outliers, developing a measurement plan is also valuable at this stage. If the plan includes expected outcomes for outliers and disability groups, this can impact what data (including people) are included, and what data and people are left out.

Finally, a word of warning. From a machine learning perspective, an obvious solution to handling a specialized sub-group not typical of the general population might be to develop a specialized model for that group. For example, a specialized speech recognition model tuned to the characteristics of people with slurred speech, or people who stutter. From an ethical perspective, solutions that handle outliers and disability groups by routing them to an alternative service require careful thinking. Individuals may not wish to self-identify as having a disability, and there may be legal protections against requiring self-declaration. Solutions that attempt to infer disability status, or infer a quality that serves as a proxy for disability status also present an ethical minefield. It may be acceptable to detect stuttered speech in order to route a speech sample to a specialized speech recognition model with higher accuracy, but using the same detection system to evaluate a job applicant could be discriminatory, unfair and potentially illegal. Any system that explicitly detects ability-related attributes of an individual may need to make these inferences visible to the user, optional, and able to be challenged when they make wrong inferences. It is crucial to involve members of the affected disability group from the outset. This can prevent wasted time and effort on paths that lead to inappropriate and unfair outcomes, inflexible systems that cannot be effectively deployed at scale, or that will be likely to face legal challenges.

Data Sourcing

When sourcing data to build a model, important considerations are:

1. Does the data include people with disabilities, especially those disabilities identified as being most impacted by this solution? For example, data about the employees of a company with poor diversity may not include anyone who is deaf, or blind. If important groups are missing, or if this information is not known, take steps to find or create such data to supplement the original data source.
2. Might the data embody bias against people with disabilities? Consider whether the

data might capture existing societal biases against people with disabilities. For example, a dataset of housing applications with decisions might reflect a historical reluctance to choose someone with a disability. When this situation is identified, raise the issue.

3. Is disability information explicitly represented in the data? If so, practitioners can use bias detection tests to check for bias, and follow up with mitigation techniques to adjust for bias before training a model (Bellamy et al., 2018).

Sometimes data are constructed by combining several data sources. Depending on the requirements of those sources, some groups of people may not have records in all sources. Be attentive to whether disability groups might fall into this category and be dropped from the combined data set. For example, when combining photograph and fingerprint biometrics, consider what should happen for individuals who do not have fingers, and how they will be represented and handled.

In Europe, GDPR regulations (European Union, 2016) give individuals the right to know what data about them is being kept, and how it is used, and to request that their data be deleted. As organizations move to limit the information they store and the ways it can be used, AI systems may often not have explicit information about disability that can be used to apply established fairness tests and corrections. By being attentive to the potential for bias in the data, documenting the diversity of the data set, and raising issues early, practitioners can avoid building solutions that will perpetuate inequalities, and identify system requirements for accommodating groups that are not represented in the data.

Data Pre-Processing

The process of cleaning and transforming data into a form suitable for machine learning has been estimated to take 80-90% of the effort of a typical data science project (Zhang, Zhang, & Yang, 2003), and the choices made at this stage can have implications for the inclusiveness of the solution.

- **Data cleaning** steps may remove outliers, presumed to be noise or measurement error, but actually representing non-typical individuals, reducing the diversity in the dataset.
- **Feature selection** may include or exclude features that convey disability status. Besides explicit disability information, other features could be impacted by disability status or the resulting societal disadvantage, providing a proxy for disability status. For example, a preference for large fonts could serve as a proxy for visual impairment, or use of video captions could be correlated with deafness. Household income, educational achievement, and many other variables can also be correlated with disability.
- **Feature engineering** involves deriving new features from the data, either through analysis or combination of existing features. For example, calculating a person's reading level or personality traits based on their writing, or calculating a ratio of days worked to days lived. In both of these examples, the derived feature will be impacted by certain disabilities.

Although accepted practice in many fields is to exclude sensitive features so as not to build a model that uses that feature, this is not necessarily the best approach for algorithmic solutions. The reality is that it can be extremely difficult to avoid including disability status in some way. When possible, including features that explicitly represent disability status allows for testing and mitigation of disability-related bias. Consulting outlier individuals and stakeholder groups identified in the problem scoping stage is valuable to provide a better understanding of the ways that disability can be reflected in the data, and the tradeoffs involved in using or excluding certain features and data values.

Preserving Privacy

People experiencing disabilities may have the most to gain from many smart systems, but are also particularly vulnerable to data abuse and misuse. The current privacy protections do not work for individuals who are outliers or different from the norm. The current response to this data abuse and misuse, by privacy efforts globally, is to de-identify the data. The notion is that if we remove our identity from the data, it can't be traced back to us and it can't be used against us. The assumption is that we will thereby retain our privacy while contributing our data to making smarter decisions about the design.

While people experiencing disabilities are particularly vulnerable to data abuse and misuse, they are often also the easiest to re-identify. If you are the only person in a neighborhood using a wheelchair, it will be easy to re-identify you. If you are the only person that receives delivery of colostomy bags in your community, it will be very easy to re-identify your purchasing data.

If de-identification is not a reliable approach to maintaining the privacy of individuals that are far from average, but data exclusion means that highly impactful decisions will be made without regard to their needs, what are potential approaches to addressing this dilemma? The primary focus has been on an ill-defined notion of privacy. When we unpack what this means to most people, it is self-determination, ownership of our own narrative, the right to know how our data is being used, and ethical treatment of our story.

To begin to address this dilemma, an International Standards Organization personal data preference standard has been proposed as an instrument for regulators to restore self-determination regarding personal data. The proposal is developed as a response to the all-or-nothing terms of service agreements which ask you to give away your private data rights in exchange for the privilege of using a service. These terms of service agreements are usually couched in legal fine print that most people could not decode even if they had the time to read them. This means that it has become a convention to simply click “I agree” without attending to the terms and the rights we have relinquished. The proposed standard will be part of an existing standard called AccessForAll or ISO/IEC 24751 (ISO/IEC, 2008). The structure of the parent standard enables matching of consumer needs and preferences with resource or service functionality. It provides a common language for describing what you need or prefer in machine-readable terms and a means for service providers or producers to describe the functions their products and services offer. This allows platforms to match diverse unmet consumer needs with the closest product or service offering. Layered on top of the standard are utilities that help consumers explore, discover and refine their understanding of their needs and preferences, for a given context and a given goal. The personal data preference part of this standard will let consumers declare what personal data they are willing to release to whom, for what purpose, what length of time and under what conditions. Services that wish to use the data would declare what data is essential for providing the service and what data is optional. This will enable a platform to support the negotiation of more reasonable terms of service. The data requirements declarations by the service provider would be transparent and auditable. The standard will be augmented with utilities that inform and guide consumers regarding the risks and implications of preference choices. Regulators in Canada and Europe plan to point to this standard when it is completed. This will hopefully wrest back some semblance of self-determination of the use of our data.

Another approach to self-determination and data that is being explored with the Platform Co-op Consortium (Platform Cooperativism Consortium, 2019) is the formation of data co-ops. In a data co-op, the data producers would both govern and share in the profit (knowledge and funds) arising from their own data. This approach is especially helpful in amassing data in previously ignored domains, such as rare illnesses, niche consumer needs or specialized hobbies. In smart cities, for example, there could be a multiplicity of data domains that could have associated data co-ops. Examples include wayfinding and traffic information, utility usage, waste management, recreation, consumer demands, to name just a few. This multiplicity of data co-ops would then collaborate to provide input into more general urban planning decisions.

Model Training and Testing

When developing a model, there are many bias testing methods available to researchers. However, when applying these techniques to fairness for people with disabilities, some limitations become evident. Firstly, group-based methods require large enough numbers of individuals in each group to allow for statistical comparison of outcomes, and secondly, they often rely on binary in-group/out-group comparisons, which can be difficult to apply to disability. This section will expand on each of these points and suggest ways to address these limitations in model testing.

When examining fairness for people with disabilities, there may be few examples in the training data. As defined by the United Nations Convention on the Rights of People with Disabilities (UN General Assembly, 2007), disability "results from the interaction between persons with impairments and attitudinal and environmental barriers that hinders their full and effective participation in society." As such, disability depends on context and comes in many forms, including physical barriers, sensory barriers, and communication barriers. One important consequence of experiencing a disability is that it can lead us to do things in a unique way, or to look or act differently. As a result, disabled people may be outliers in the data, or align with one of many very different sub-groups.

Today's methods for bias testing tend to split individuals into members of a protected group, and 'others'. However, disability status has many dimensions, varies in intensity and impact, and often changes over time. Furthermore, people are often reluctant to reveal a disability. Binarized approaches that combine many different people into a broad 'disabled' category will fail to detect patterns of bias that apply, differently and distinctively, against specific groups within the disability umbrella. For example, an inaccessible online testing Web site will not disadvantage wheelchair users, but those who rely on assistive technologies or keyboard-only control methods to access the Web may be unable to complete the tests. More sensitive analysis methods not based on binary classifications are needed to address these challenges. Other protected attributes like race and gender that have traditionally been examined with binary fairness metrics are also far more complex and nuanced in reality (Keyes, 2018) (Hamidi, Scheuerman, & Branham, 2018b), and new approaches suitable for examining disability-related bias would support a more sophisticated examination of these attributes too.

To test for fairness, an audit based on identified test cases and expected outcomes is valuable. These can be developed with the outlier individuals and key stakeholder groups identified at the outset. For models that interpret humans (speech, language, gesture, facial expression) where algorithmic fairness is important, the goal is to develop a method that works well for as many groups as possible (e.g. speech recognition for deaf speakers), and to document the limitations of the model. For models that allocate people to groups (e.g. job candidate, loan applicant), allocative fairness is important. In selecting a measure for allocative fairness, we argue that an individual fairness approach rather than a group fairness approach is preferable. While group fairness seeks to equalize a measure across groups, individual fairness aims for 'similar' individuals to receive similar outcomes. For example, in deciding whether to grant loan applications, even if the presence of a disability is statistically correlated with unemployment over the whole dataset, it would still be unfair to treat an employed person with a disability the same as the unemployed group, simply because of their disability. Individual fairness aligns better with the societal notion of fairness, and legal mandates against discrimination.

Deployment in Real Applications

In this stage, the trained model is incorporated into an application, typically with an interface for people to use, or an API to connect to. Testing with diverse users, especially outliers, is essential. Understanding how different people will perceive and use an AI-based system is also important, for example, to see if certain people are more likely than others to ascribe trust to an AI-based system, or be more likely to feel insulted by an AI system's terse replies or lack of context.

As a matter of course, quality assurance should include testing by people with disabilities, covering as broad a set of disability groups as possible. This would include both testing the user interface of the system itself to ensure it is accessible, and the system's performance on diverse data inputs. The test process should deliberately include outlier individuals to test the limits of the system and the mechanisms for addressing system failure.

Because disability manifests in such diverse ways, there will be situations where the application is presented with an individual quite unlike those in the training data. For example an automated telephone help system may have difficulty interpreting an individual with a speech impairment, especially if they are not speaking in their native language. Developers can avoid discrimination by providing an alternative to using AI, for example by supporting typed input in addition to speech. Users should have the ability to opt out of AI-based interpretation.

Disability may also impact the accuracy of inputs to an AI-based system. An example is a video-based personality analysis concluding that an autistic interviewee is untrustworthy because they did not make eye contact with the interviewer, and then feeding that into an applicant selection model. For people with disabilities, it is essential to have the opportunity to inspect and correct the data used to make decisions about them.

Equally important is the ability to query and challenge AI decisions, receiving some form of explanation of the factors that most impacted the decision. If these factors were affected by a person's disability, the decision may be discriminatory. Any AI-based system that makes decisions affecting people should include both an opportunity to dispute a decision, and provide a manual override for outlier individuals where the model is unreliable.

As many AI systems by their very nature learn and thus modify their behavior over time, ongoing mechanisms to monitor for fairness to people with disabilities should be incorporated. These can include ongoing auditing and reviews of performance as well as periodic explicit testing to verify that changes in the system's operation aimed at improving performance do not introduce disparities in how decisions are made for specific subpopulations. This is crucial to ensure that as a system gets better for people overall it doesn't unfairly get worse for some.

Design Approaches

In several of the stages of AI development described above, and particularly in the problem scoping and testing/deployment phases, we have encouraged AI/ML engineers to seek the ongoing involvement of people with disabilities. For AL/ML practitioners, this may seem like a daunting task, with questions ranging from how to find diverse users, how to ethically and respectfully engage them, and by what methods one can reliably incorporate their feedback to improve systems.

The field of Human-Computer Interaction has long contemplated these questions, and has developed a number of design philosophies, methodologies, and methods to guide the practice. Some of these pertain specifically to engaging people with disabilities, including Universal Design (Story, Mueller, & Mace, 1998), Ability-Based Design (Wobbrock, Kane, Gajos, Harada, & Froehlich, 2011), Design for User Empowerment (Ladner, 2015), and Design for Social Accessibility (Shinohara, Wobbrock, & Pratt, 2018). In this section, we endeavor to provide a brief overview of just three potential approaches that AL/ML developers might choose as they seek to integrate users into their process. Our purpose, then, is not to provide a comprehensive review of all or even a few methodologies, but rather to offer links into the literature for those who want to learn more about these approaches, or perhaps seek out a collaborator with such expertise.

Below, we overview three distinct approaches to human-centered design: Inclusive Design, Participatory Design, and Value-Sensitive Design. Each of these has developed from different intellectual traditions, and therefore varies in the degree to which they explicitly include people with disabilities in their theoretical frameworks. People with disabilities are often excluded from design processes, and designs rarely anticipate end-users' needs to appropriate and adapt designs (Derboven, Geerts, & De Grooff, 2016). We therefore will begin here by introducing some basic rationale for why it is important to include people with disabilities directly in software development efforts.

Firstly, the opportunity to fully participate in society is every person's right, and digital inclusion is fundamental to those opportunities today. All of us will very likely experience disability at some stage in our lives, and our technology must be robust enough to accommodate the diversity of the human experience, especially if it is used in critical decision-making areas. This cannot happen unless this diversity is considered from the outset, hence the disability rights movement's mantra of "nothing about us, without us."

Including people with disabilities may bring compelling new design ideas and ultimately expand the potential user base of the product. People with disabilities (including seniors) have been described as the original life hackers and personal innovators (Harley & Fitzpatrick, 2012) (Storni, 2010), as they often have to find creative workarounds and innovate new technologies in a world that is not built to their specifications. Second, people with disabilities can be seen as presenting valuable diverse cases which should be brought from the "edge" and into the "center" of our design thinking. In her foundational paper on Feminism in Human Computer Interaction, Bardzell advocated to study not only the conceptual "center" of a distribution of users, but also the edge cases (Bardzell, 2010). She argued that design often functions with a default "user" in the designers' minds, and that default user is often male, white, educated, and non-disabled. In Bardzell's analysis, accommodating the edge cases is a way both to broaden the audience (or market) for a design, and to strengthen the design against unanticipated changes in users, usage, or contexts of use. These ideas are echoed by other researchers (Krischkowsky et al., 2015) (Muller et al., 2016) (Tscheligi et al., 2014). One canonical example of how designs made with and for people with disabilities can actually improve the user experience of everyone (an example of "universal design"), is the curb cut. Curb cuts – the ramps that allow people in wheelchairs to transition from public sidewalks to cross the street – also serve parents with prams, workers with heavy wheeled loads, and pedestrians on scooters. Both Downey and Jacobs (Downey, 2008) (Jacobs, 1999) have advocated for electronic curb cuts; one example of such a feature is the zooming capability of the browser, which supports easier reading for people with low vision or people who are far away from the screen.

In practice, the best way of centering marginalized perspectives will require that we include people with disabilities in our core design practices. Fortunately, there is a rich history of work on design that centers users at the margins. In particular, we believe the theoretical approaches of Inclusive Design (specifically as it evolved in Canada), Participatory Design and Value Based Design are particularly valuable when designing for – and with – people with disabilities.

Inclusive Design

The practice and theoretical framing of inclusive design, that emerged and evolved in Canada and with global partners since the emergence of the Web, takes advantage of the affordances or characteristics of digital systems (Pullin, Treviranus, Patel, & Higginbotham, 2017) (Treviranus, 2016). In contrast to related universal design theories, that emerged from architectural and industrial design fields, the Canadian inclusive design practice aims to use the mutability and connectivity of networked digital systems to achieve one-size-fits-one designs within an integrated system, thereby increasing the adaptability and longevity of the system as a whole (Lewis & Treviranus, 2013). Rather than specifying design criteria, or accessibility checklists, the theory specifies a process or mindset, called the "three dimensions of inclusive design."

1. Recognize that everyone is unique, and strive for a design that is able to match this uniqueness in an integrated system. Support self-awareness of this uniqueness (use data to make people 'smarter' about themselves, not just machines smarter).
2. Create an inclusive co-design process. The most valuable co-designers are individuals that can't use or have difficulty using the current designs. Continuously ask whose perspective is missing from the decision making "table" and how can they help make the "table" more inclusive.
3. Recognize that all design operates within a complex adaptive system of systems. Be cognizant of the entangled impact and friction points. Strive for designs that are beneficial to this larger system of systems.

This practice of inclusive design critiques and counters reliance on probability and population based statistics, pointing out the risks of basing critical decisions solely on the majority or statistical average (Treviranus, 2014). With respect to fairness in AI, people with disabilities are most disadvantaged by population-based AI decisions. The only common defining characteristic of disability is difference from the norm. If you are not like the norm, probability predictions based on population data are wrong. Even if there is full representation and all human bias is removed, statistically-based decisions and predictions will be biased against small minorities and outliers. Current automated decisions focus on the good of the majority, causing greater disparity between the majority and smaller minorities. Inclusive design practitioners are investigating learning models that do not give advantage to being like the majority, causing learning models to attend to the full diversity of requirements (Treviranus, 2019). This is hypothesized to also support better context shifting, adaptability and detection of weak signals.

Given that data is from the past, optimization using past data will not achieve the culture shift inclusive design practitioners hope to achieve. Hence inclusive design practitioners are combining existing data, data from alternative scenarios, and modelled or simulated data to assist in decision making. Storytelling, bottom-up personalized data, or small (n=1), thick (in context) data is also employed to overcome the bias toward numerical data (Clark, 2018) (Pullin et al., 2017).

Participatory Design

Participatory Design (PD) emphasizes the active role of people who will be affected by a technology, as co-designers of that technology. There are many methods and rationales for these approaches, which can be found in (Muller & Druin, 2012), among other references. Some PD methods were originally proposed as "equal opportunity" practices, e.g., (Kuhn & Muller, 1993), because the methods involved low-technology or "lo-fi" prototyping practices that did not require extensive computer knowledge to contribute to the design. However, the needs of people with disabilities were generally not considered during the early days of PD.

This omission has now been partially rectified. Börjesson and colleagues (Börjesson, Barendregt, Eriksson, & Torgersson, 2015) recently published an overview of theory and methods for work with developmentally diverse children. Katan et al. (2015) used interactive machine learning in participatory workshops with people with disabilities (Katan, Grierson, & Fiebrink, 2015).

Some of these methods amount to merely asking people about their needs, e.g., (Holbø, Bøthun, & Dahl, 2013) (Krishnaswamy, 2017). However, other approaches involve bringing people with disabilities (and sometimes their informal carers) into the design process as active co-designers, e.g., (Gomez Torres, Parmar, Aggarwal, Mansur, & Guthrie, 2019) (Hamidi et al., 2016) (Lee & Riek, 2018) (McGrenere et al., 2003) (Sitbon & Farhin, 2017) (Wilde & Marti, 2018) (Williams et al., 2018). In general, the methods that include direct participation by people with disabilities in design activities are more powerful, and tend to include deeper understandings than are possible through the less engaged survey methods.

We note that this is an active research area, with discussion of needs that are not yet met, e.g., (Holone & Herstad, 2013) (Oswal, 2014), and many opportunities to improve and innovate the participatory methods. For people who are new to PD, we suggest beginning with an orientation to the diversity of well-tested methods, e.g., (Muller & Druin, 2012), followed by a "deeper dive" into methods that have been used with particular populations and/or with particular challenges.

Value-Sensitive Design

Participatory design originated primarily from the workplace democracy movement in Scandinavia, e.g., (Bjerknes, Ehn, & Kyng, 1987), and then developed in many directions. One of the core assumptions of the workplace applications was a division of labor among workers and managers. In that context, PD methods were seen as ways to reduce power differences during the two-party design process, by facilitating the voice of the workers in relation to management. These assumptions have tended to carry through into work with people with disabilities, in which the two parties are reconceived as people with disabilities and designers, or people with disabilities and providers, with designers as mediators.

Value Sensitive Design (VSD) offers a broader perspective regarding the stakeholders in design (Friedman, Hendry, & Borning, 2017). For this paper, VSD crucially expands concepts of stakeholders into direct stakeholders (people who have contact with a design or a technology, including users, designers, providers) and indirect stakeholders (people who are affected by the design or technology, even if they do not have direct contact with it). For people with disabilities, there are often multiple stakeholders in complex relationships, e.g., (Zolyomi, Ross, Bhattacharya, Milne, & Munson, 2018).

While there are disagreements about the details of a value-centric approach, e.g., (Borning & Muller, 2012) (Le Dantec, Poole, & Wyche, 2009) (Muller & Liao, 2017), there is consensus that values matter, and that values can be formative for designs. There may be values differences among people with disabilities, carers, and medical professionals, e.g, (Draper et al., 2014) (Felzmann, Beyan, Ryan, & Beyan, 2016), and therefore an explicit and focused values inquiry may be helpful in "satisficing" these complex assumptions, views, and needs (Cheon & Su, 2016). While VSD has tended to have fewer well-defined methods, Friedman et al. (2017) recently published a survey of values-centric methods (Friedman et al., 2017).

Conclusion

We have outlined a number of situations in which AI solutions could be disadvantageous for people with disabilities if researchers and practitioners fail to take necessary steps. In many existing situations, non-AI solutions are already discriminatory, and introducing AI runs the risk of simply perpetuating and replicating these flaws. For example, people with disabilities may already face discrimination in hiring opportunities. With AI-driven hiring systems, models that recognize good candidates by matching to the existing workforce will perpetuate that status quo. In education, an AI system that draws inferences based on a student's online interactions might misinterpret speed for competency, if the student is using assistive technologies. In public safety, AI systems might misinterpret a person with a cognitive disability as a potential threat. In AI systems for healthcare, where speech characteristics can be used to diagnose cognitive impairments, a person with a speech impediment can be misdiagnosed.

To avoid such erroneous conclusions and potentially damaging outcomes, a number of steps are proposed. AI systems should be prioritized for fairness review and ongoing monitoring, based on their potential impact on the user in their broader context of use. They should offer opportunities to redress errors, and for users and those impacted to raise fairness concerns. People with disabilities should be included when sourcing data to build models. Such "outlier" data - the edge cases - will create a more inclusive and robust system. From the perspective of people with disabilities, there can be privacy concerns with self-identification, but there can be risk of exclusion from the data models if users opt not to participate or disclose. There are methods provided to increase participation while protecting user privacy, such as the personal data preferences standard. In deploying the AI application, it is critical to test the UI and system preferences with outlier individuals. Users should be able to pursue workarounds, and ultimately override the system where models may be unreliable.

AI has been shown to help improve the lives of people with disabilities in a number of different environments whether it be navigating a city, re-ordering prescriptions at a local pharmacy through a telephone or text service, or facilitating public safety. Almost everyone in the greater community is directly connected with someone with a disability whether it be a family member, a colleague, a friend, or a neighbor. While AI technology has significantly improved the lives of those in the disabled community, there are always ways in which we can continue to advocate for fairness and equality and challenge the status quo.

When creating AI that is designed to help the community, we must take into consideration a disabled person user approach. The AI designed should consider disabled people as a focal point. To borrow upon the concept driven by Eric Ries in The Lean Startup (Ries, 2011), we need to deploy minimum viable products (MVPs) that are not perfected but rather improved upon by the user involved.

In a sense, what is needed is an incremental and algorithmic approach that continues to challenge the status quo and strives to improve the standardization of fairness and equality. This should be from a multi-industrial approach with key players in different industries.

This paper suggests challenging every day practices that may prove to inhibit people with disabilities, and is a starting point to bring awareness for the need for equality. It is important to remember that promoting this goal is a process. Success will require a series of incremental steps of further learning, thought provoking peer discussion, and changes at the local and municipal level. Only when these incremental changes are met will we drive sustainable outcomes for people with disabilities using AI systems.

Acknowledgments

The authors thank all participants of the IBM-MIT AI Week 2018 Workshop on AI Fairness for People with Disabilities for their contributions to the discussions leading to this article. Special thanks to Peter Fay, Phill Jenkins, Audrey Hasegawa and Morgan Foreman.

References

Ameri, M., Schur, L., Adya, M., Bentley, F. S., McKay, P., & Kruse, D. (2018, mar). The Disability Employment Puzzle: A Field Experiment on Employer Hiring Behavior. *ILR Review*, *71*(2), 329–364. Retrieved from http://journals.sagepub.com/doi/10.1177/0019793917717474 doi: 10.1177/0019793917717474

Bardzell, S. (2010). Feminist HCI: Taking Stock and Outlining an Agenda for Design. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1301–1310). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1753326.1753521 doi: 10.1145/1753326.1753521

Barnett, S., McKee, M., Smith, S. R., & Pearson, T. A. (2011, mar). Deaf sign language users, health inequities, and public health: opportunity for social justice. *Preventing chronic disease*, *8*(2), A45. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21324259http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3073438

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., . . . Zhang, Y. (2018, oct). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. Retrieved from http://arxiv.org/abs/1810.01943

Bhutani, A., & Wadhwani, P. (2019). *Artificial Intelligence (AI) in Education Market Size worth $6bn by 2024.* Global Market Insights. Retrieved from https://www.gminsights.com/pressrelease/artificial-intelligence-ai-in-education-market

Bioss. (2019). *The Bioss AI Protocol.* Retrieved 2019-08-30, from http://www.bioss.com/ai/

Bird, S., Hutchinson, B., Kenthapadi, K., Kiciman, E., & Mitchell, M. (2019). Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned. In *Companion proceedings of the 2019 world wide web conference on - www '19* (pp. 1297–1298). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=3308560.3320086 doi: 10.1145/3308560.3320086

Bjerknes, G., Ehn, P., & Kyng, M. (1987). *Computers and Democracy - A Scandinavian Challenge.* Avebury.

Börjesson, P., Barendregt, W., Eriksson, E., & Torgersson, O. (2015). Designing Technology for and with Developmentally Diverse Children: A Systematic Literature Review. In *Proceedings of the 14th international conference on interaction design and children* (pp. 79–88). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2771839.2771848 doi: 10.1145/2771839.2771848

Borning, A., & Muller, M. (2012). Next Steps for Value Sensitive Design. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1125–

1134). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2207676.2208560 doi: 10.1145/2207676.2208560

Branham, S. M., Abdolrahmani, A., Easley, W., Scheuerman, M., Ronquillo, E., & Hurst, A. (2017). ”Is Someone There? Do They Have a Gun”. In *Proceedings of the 19th international acm sigaccess conference on computers and accessibility - assets ’17* (pp. 260–269). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=3132525.3132534 doi: 10.1145/3132525.3132534

Brewer, R. N., & Kameswaran, V. (2018). Understanding the Power of Control in Autonomous Vehicles for People with Vision Impairment. In *Proceedings of the 20th international acm sigaccess conference on computers and accessibility - assets ’18* (pp. 185–197). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=3234695.3236347 doi: 10.1145/3234695.3236347

Bruckner, M. A. (2018). The Promise and Perils of Algorithmic Lenders’ Use of Big Data. *Chicago-Kent Law Review*, *93*. Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/chknt93{&}id=15{&}div={&}collection=

Bureau of Labor Statistics, U. D. o. L. (2019). *PERSONS WITH A DISABILITY: LABOR FORCE CHARACTERISTICS — 2018.* Retrieved from https://www.bls.gov/news.release/pdf/disabl.pdf

Burgstahler, S. (2015, dec). Opening Doors or Slamming Them Shut? Online Learning Practices and Students with Disabilities. *Social Inclusion*, *3*(6), 69. Retrieved from http://www.cogitatiopress.com/ojs/index.php/socialinclusion/article/view/420 doi: 10.17645/si.v3i6.420

Chander, A. (2017). The Racist Algorithm? *Michigan Law Review*, *115*(6), 1023. Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/mlr115{&}id=1081{&}div={&}collection=

Cheon, E., & Su, N. M. (2016). Integrating roboticist values into a Value Sensitive Design framework for humanoid robots. *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. doi: 10.1109/HRI.2016.7451775

Cinquin, P.-A., Guitton, P., & Sauzéon, H. (2019, mar). Online e-learning and cognitive disabilities: A systematic review. *Computers & Education*, *130*, 152–167. Retrieved from https://www.sciencedirect.com/science/article/pii/S0360131518303178 doi: 10.1016/J.COMPEDU.2018.12.004

Clark, C. B. (2018, jun). *Many more stories: Co-design and creative communities.* OCAD University. Retrieved from http://openresearch.ocadu.ca/id/eprint/2079/

Costello, K. (2019). *Gartner Survey Shows 37 Percent of Organizations Have Implemented AI in Some Form.* Stamford, Conn.: Gartner. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2019-01-21-gartner-survey-shows-37-percent-of-organizations-have

Cutler, A., Pribik, M., & Humphrey, L. (2019). *Everyday Ethics for Artificial Intelligence* (Tech. Rep.). IBM. Retrieved from https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women - Reuters.* Retrieved 2019-08-30, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Derboven, J., Geerts, D., & De Grooff, D. (2016). The Tactics of Everyday Practice: A Semiotic Approach to Appropriation. *Interaction Design and Architecture(s)*, *29*(29), 99–120. Retrieved from https://lirias.kuleuven.be/389602?limo=0

Downey, G. (2008). *Closed captioning: Sub-*

titling, stenography and the digital convergemce of text with television. Johns Hopkins University Press.

Draper, H., Sorell, T., Bedaf, S., Syrdal, D. S., Gutierrez-Ruiz, C., Duclos, A., & Amirabdollahian, F. (2014, oct). Ethical Dimensions of Human-Robot Interactions in the Care of Older People: Insights from 21 Focus Groups Convened in the UK, France and the Netherlands. In *International conference on social robotics* (LNCS 8755 ed., pp. 135–145). Springer. Retrieved from http://link.springer.com/10.1007/978-3-319-11973-1{_}14 doi: 10.1007/978-3-319-11973-1_14

Dudley-Marling, C., & Burns, M. B. (2014). Two Perspectives on Inclusion in the United States, Global Education Review, 2014. *Global Education Review*, *1*(1), 14–31. Retrieved from https://eric.ed.gov/?id=EJ1055208

European Union. (2016). *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 (General Da.*

Faggella, D. (2019, apr). *Machine Learning in Human Resources – Applications and Trends.*

Faucett, H. A., Ringland, K. E., Cullen, A. L. L., & Hayes, G. R. (2017, oct). (In)Visibility in Disability and Assistive Technology. *ACM Trans. Access. Comput.*, *10*(4), 14:1—-14:17. Retrieved from http://doi.acm.org/10.1145/3132040 doi: 10.1145/3132040

Felzmann, H., Beyan, T., Ryan, M., & Beyan, O. (2016, jan). Implementing an Ethical Approach to Big Data Analytics in Assistive Robotics for Elderly with Dementia. *SIGCAS Comput. Soc.*, *45*(3), 280–286. Retrieved from http://doi.acm.org/10.1145/2874239.2874279 doi: 10.1145/2874239.2874279

Friedman, B., Hendry, D. G., & Borning, A. (2017). A Survey of Value Sensitive Design Methods. *Foundations and Trends* R *in Human–Computer Interaction*, *11*(2), 63–125. Retrieved from http://dx.doi.org/10.1561/1100000015 doi: 10.1561/1100000015

Fruchterman, J., & Mellea, J. (2018). *Expanding Employment Success for People with Disabilities — Benetech — Software for Social Good* (Tech. Rep.). benetech. Retrieved from https://benetech.org/about/resources/expanding-employment-success-for-people-with-disabilities-2/

Gajos, K. Z., Wobbrock, J. O., & Weld, D. S. (2007). Automatically Generating User Interfaces Adapted to Users' Motor and Vision Capabilities. In *Proceedings of the 20th annual acm symposium on user interface software and technology* (pp. 231–240). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1294211.1294253 doi: 10.1145/1294211.1294253

Givens, A. R. (2019). *Institute Announces New Project on Algorithmic Fairness for People with Disabilities.* Institute for Technology Law and Policy, Georgetown Law.

Glickman, N. (2007). *Do You Hear Voices? Problems in Assessment of Mental Status in Deaf Persons With Severe Language Deprivation* (Vol. 12). Oxford University Press. Retrieved from https://www.jstor.org/stable/42658866 doi: 10.2307/42658866

Gomez Torres, I., Parmar, G., Aggarwal, S., Mansur, N., & Guthrie, A. (2019). Affordable Smart Wheelchair. In *Extended abstracts of the 2019 chi conference on human factors in computing systems* (pp. SRC07:1—-SRC07:6). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3290607.3308463 doi: 10.1145/3290607.3308463

Guo, A., Kamar, E., Vaughan, J. W., Wallach, H., & Morris, M. R. (2019, jul). *Toward Fairness in AI for People with Disabilities: A Research Roadmap.* Retrieved from http://arxiv.org/abs/1907.02227

Haigh, K. Z., & Yanco, H. A. (2002). *Automation as Caregiver: A Survey of Issues* *and Tchnologies* (Tech. Rep.). AAAI.

Hamidi, F., Müller, C., Baljko, M., Schorch, M., Lewkowicz, M., & Stangl, A. (2016). Engaging with Users and Stakeholders: The Emotional and the Personal. In *Proceedings of the 19th international conference on supporting group work* (pp. 453–456). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2957276.2996292 doi: 10.1145/2957276.2996292

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018a). Gender Recognition or Gender Reductionism? In *Proceedings of the 2018 chi conference on human factors in computing systems - chi '18* (pp. 1–13). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=3173574.3173582 doi: 10.1145/3173574.3173582

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018b). Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 8:1—-8:13). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3173574.3173582 doi: 10.1145/3173574.3173582

Hankerson, D., Marshall, A. R., Booker, J., El Mimouni, H., Walker, I., & Rode, J. A. (2016). Does Technology Have Race? In *Proceedings of the 2016 chi conference extended abstracts on human factors in computing systems - chi ea '16* (pp. 473–486). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=2851581.2892578 doi: 10.1145/2851581.2892578

Harley, D., & Fitzpatrick, G. (2012). Appropriation of social networking by older people: two case studies. In *Ecsw 2011 workshop on fostering social interactions in the ageing society: Artefacts - methodologies - research paradigms.* Aarhus, Denmark.

Harrell, E. (2017). *Crime Against Persons with Disabilities, 2009-2015 - Statistical Tables* (Tech. Rep.). Bureau of Justice Statistics. Retrieved from http://www.bjs.gov/index.cfm?ty=pbdetail{&}iid=5986

Holbø, K., Bøthun, S., & Dahl, Y. (2013). Safe Walking Technology for People with Dementia: What Do They Want? In *Proceedings of the 15th international acm sigaccess conference on computers and accessibility* (pp. 21:1—-21:8). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2513383.2513434 doi: 10.1145/2513383.2513434

Holone, H., & Herstad, J. (2013). Three Tensions in Participatory Design for Inclusion. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2903–2906). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2470654.2481401 doi: 10.1145/2470654.2481401

Hurley, M., & Adebayo, J. (2016). Credit Scoring in the Era of Big Data. *Yale Journal of Law and Technology*, *18*. Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/yjolt18{&}id=148{&}div={&}collection=

IEEE, G. I. o. E. o. A., & Systems, I. (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (First edit ed.). IEEE. Retrieved from https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/{%}0Aautonomous-systems.html

Iezzoni, L. I. (2011, oct). Eliminating Health And Health Care Disparities Among The Growing Population Of People With Disabilities. *Health Affairs*, *30*(10), 1947–1954. Retrieved from http://www.healthaffairs.org/doi/10.1377/hlthaff.2011.0613 doi: 10.1377/hlthaff.2011.0613

ISO/IEC. (2008). *Information technology — Individualized adaptability and accessibility in e-learning, education and training — Part 1: Framework and reference model.* Author.

Jacobs, S. (1999). *Section 255 of the Telecommunications Act of 1996: Fueling the Creation of New Elec-*

tronic Curbcuts. The Center for an Accessible Society. Retrieved from http://www.accessiblesociety.org/topics/technology/eleccurbcut.htm

Janssen, M., & Kuk, G. (2016, jul). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly*, *33*(3), 371–377. Retrieved from https://www.sciencedirect.com/science/article/pii/S0740624X16301599 doi: 10.1016/J.GIQ.2016.08.011

Kanellopoulos, Y. (2018, jul). A Model for Evaluating Algorithmic Systems Accountability. Retrieved from http://arxiv.org/abs/1807.06083

Katan, S., Grierson, M., & Fiebrink, R. (2015). Using Interactive Machine Learning to Support Interface Development Through Workshops with Disabled People. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 251–254). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2702123.2702474 doi: 10.1145/2702123.2702474

Keyes, O. (2018, nov). The Misgendering Machines. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–22. Retrieved from http://dl.acm.org/citation.cfm?doid=3290265.3274357 doi: 10.1145/3274357

Koene, A., Dowthwaite, L., & Seth, S. (2018). IEEE P7003TM standard for algorithmic bias considerations. In *Proceedings of the international workshop on software fairness - fairware '18* (pp. 38–41). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=3194770.3194773 doi: 10.1145/3194770.3194773

Koene, A., Smith, A. L., Egawa, T., Mandalh, S., & Hatada, Y. (2018). IEEE P70xx, Establishing Standards for Ethical Technology. *Proceedings of KDD, ExCeL London UK, August, 2018 (KDD'18).*

Krahn, G. L., & Fox, M. H. (2014, sep). Health disparities of adults with intellectual disabilities: what do we know? What do we do? *Journal of applied research in intellectual disabilities : JARID*, *27*(5), 431–46. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23913632http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4475843 doi: 10.1111/jar.12067

Krahn, G. L., Hammond, L., & Turner, A. (2006, jan). A cascade of disparities: Health and health care access for people with intellectual disabilities. *Mental Retardation and Developmental Disabilities Research Reviews*, *12*(1), 70–82. Retrieved from http://doi.wiley.com/10.1002/mrdd.20098 doi: 10.1002/mrdd.20098

Krahn, G. L., Walker, D. K., & Correa-De-Araujo, R. (2015, apr). Persons with disabilities as an unrecognized health disparity population. *American journal of public health*, *105 Suppl*(Suppl 2), S198–206. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25689212http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4355692 doi: 10.2105/AJPH.2014.302182

Krischkowsky, A., Tscheligi, M., Neureiter, K., Muller, M., Polli, A. M., & Verdezoto, N. (2015). Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design. In *Proceedings of the 14th european conference on computer supported cooperative work.*

Krishnaswamy, K. (2017). Participatory Design: Repositioning, Transferring, and Personal Care Robots. In *Proceedings of the companion of the 2017 acm/ieee international conference on human-robot interaction* (pp. 351–352). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3029798.3034815 doi: 10.1145/3029798.3034815

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable Algorithms. *University of Pennsylvania Law Review*, *165*. Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/pnlr165{&}id=648{&}div={&}collection=

Kuhn, S., & Muller, M. (1993). Participatory Design. *Communications of the ACM*, *36*(6).

K.W. v. Armstrong, No. 14-35296 (9th Cir. 2015) :: Justia. (2015). Retrieved from https://law.justia.com/cases/federal/appellate-courts/ca9/14-35296/14-35296-2015-06-05.html

Ladner, R. (2015). Design for user empowerment. *Interactions*, *XXII*(2). Retrieved from http://interactions.acm.org/archive/view/march-april-2015/design-for-user-empowerment

Le Dantec, C. A., Poole, E. S., & Wyche, S. P. (2009). Values As Lived Experience: Evolving Value Sensitive Design in Support of Value Discovery. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1141–1150). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1518701.1518875 doi: 10.1145/1518701.1518875

Lee, H. R., & Riek, L. D. (2018, may). Reframing Assistive Robots to Promote Successful Aging. *ACM Trans. Hum.-Robot Interact.*, *7*(1), 11:1—-11:23. Retrieved from http://doi.acm.org/10.1145/3203303 doi: 10.1145/3203303

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018, dec). Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology*, *31*(4), 611–627. Retrieved from http://link.springer.com/10.1007/s13347-017-0279-x doi: 10.1007/s13347-017-0279-x

Lewis, C. (2019, jun). Implications of Developments in Machine Learning for People with Cognitive Disabilities. *SIGACCESS Newsletter, Issue 124*. Retrieved from http://www.sigaccess.org/newsletter/2019-06/lewis.html

Lewis, C., & Treviranus, J. (2013, sep). Public Policy and the Global Public Inclusive Infrastructure Project. *interactions*, *20*(5), 62–66. Retrieved from http://doi.acm.org/10.1145/2510123 doi: 10.1145/2510123

Lohia, P. K., Natesan Ramamurthy, K., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2019, may). Bias Mitigation Post-processing for Individual and Group Fairness. In *Icassp 2019 - 2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 2847–2851). IEEE. Retrieved from https://ieeexplore.ieee.org/document/8682620/ doi: 10.1109/ICASSP.2019.8682620

McCarthy, O. J. (2019). *AI & Global Governance: Turning the Tide on Crime with Predictive Policing - United Nations University Centre for Policy Research.* United Nations University Centre for Policy Research. Retrieved from https://cpr.unu.edu/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html

McGrenere, J., Davies, R., Findlater, L., Graf, P., Klawe, M., Moffatt, K., ... Yang, S. (2003). Insights from the Aphasia Project: Designing Technology for and with People Who Have Aphasia. In *Proceedings of the 2003 conference on universal usability* (pp. 112–118). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/957205.957225 doi: 10.1145/957205.957225

Morris, R. R., Kirschbaum, C. R., & Picard, R. W. (2010). Broadening Accessibility Through Special Interests: A New Approach for Software Customization. In *Proceedings of the 12th international acm sigaccess conference on computers and accessibility* (pp. 171–178). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1878803.1878834 doi: 10.1145/1878803.1878834

Muller, M., & Druin, A. (2012). Participatory design: The third space of HCI. In J. Jacko (Ed.), *Human computer interaction handbook* (3rd ed., pp. 1125–1154). CRC Press.

Muller, M., & Liao, V. (2017). *Using participatory design fictions to explore ethics and values for robots and agents.*

Muller, M., Neureiter, K., Krischkowsky, A., Tscheligi, M., Al Zubaidi-Polli, A. M., & Tscheligi, M. (2016). Collaborative Appropriation: How Couples, Teams, Groups and Communities Adapt and Adopt Technologies. In *Proceedings of the 19th acm conference on computer supported cooperative work and social computing companion - cscw '16 companion* (pp. 473–480). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=2818052.2855508 doi: 10.1145/2818052.2855508

Obiakor, F. E., Harris, M., Mutua, K., Rotatori, A., & Algozzine, B. (2012). Making Inclusion Work in General Education Classrooms. *Education and Treatment of Children*, *35*(3), 477–490. doi: 10.1353/etc.2012.0020

Oswal, S. K. (2014, may). Participatory Design: Barriers and Possibilities. *Commun. Des. Q. Rev*, *2*(3), 14–19. Retrieved from http://doi.acm.org/10.1145/2644448.2644452 doi: 10.1145/2644448.2644452

Platform Cooperativism Consortium. (2019). *What Is a Platform Co-op?* Retrieved 2019-08-30, from https://platform.coop/

Policing, P. T. F. o. s. C. (2015). *Final Report of the President's Task Force on 21st Century Policing.* Washington, DC: Office of Community Oriented Policing Services. Retrieved from https://ric-zai-inc.com/Publications/cops-p311-pub.pdf

Pollard, R. (1994). Public mental health service and diagnostic trends regarding individuals who are deaf or hard of hearing. *Rehabilitation Psychology*, *39*(3), 147–160.

Popovich, P. M., Scherbaum, C. A., Scherbaum, K. L., & Polinko, N. (2003, mar). The Assessment of Attitudes Toward Individuals With Disabilities in the Workplace. *The Journal of Psychology*, *137*(2), 163–177. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/00223980309600606 doi: 10.1080/00223980309600606

Pradhan, A., Mehta, K., & Findlater, L. (2018). "Accessibility Came by Accident". In *Proceedings of the 2018 chi conference on human factors in computing systems - chi '18* (pp. 1–13). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=3173574.3174033 doi: 10.1145/3173574.3174033

Pullin, G., Treviranus, J., Patel, R., & Higginbotham, J. (2017). Designing interaction, voice, and inclusion in AAC research. *Augmentative and Alternative Communication*, *33*(3), 139–148. Retrieved from https://doi.org/10.1080/07434618.2017.1342690 doi: 10.1080/07434618.2017.1342690

Ries, E. (2011). *The lean startup : how today's entrepreneurs use continuous innovation to create radically successful businesses.* Retrieved from https://books.google.com/books?hl=en{&}lr={&}id=r9x-OXdzpPcC{&}oi=fnd{&}pg=PA15{&}dq=+The+Lean+Startup{&}ots=0s-bDboHfV{&}sig=jVn-Rw46-A2D4H{_}9HnK7gTi5MCg{#}v=onepage{&}q=TheLeanStartup{&}f=false

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the conference on fairness, accountability, and transparency - fat* '19* (pp. 59–68). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=3287560.3287598 doi: 10.1145/3287560.3287598

Shaffer, I. R., & Rogan, I. (2018). Exploring the Performance of Facial Expression Recognition Technologies on Deaf Adults and Their Children. In *Proceedings of the 20th international acm sigaccess conference on computers and accessibility - assets '18* (pp. 474–476). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=3234695.3240986 doi: 10.1145/3234695.3240986

Shaheen, N. L., & Lazar, J. (2018). K–12 Technology Accessibility: The Message from State Government. *Journal of Special Education Technology*, *33*(2), 83–97.

Shinohara, K., Wobbrock, J. O., & Pratt, W. (2018). Incorporating Social Factors in

Accessible Design. In *Proceedings of the 20th international acm sigaccess conference on computers and accessibility - assets '18* (pp. 149–160). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=3234695.3236346 doi: 10.1145/3234695.3236346

Sitbon, L., & Farhin, S. (2017). Co-designing Interactive Applications with Adults with Intellectual Disability: A Case Study. In *Proceedings of the 29th australian conference on computer-human interaction* (pp. 487–491). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3152771.3156163 doi: 10.1145/3152771.3156163

Storni, C. (2010). Multiple Forms of Appropriation in Self-Monitoring Technology: Reflections on the Role of Evaluation in Future Self-Care. *International Journal of Human–Computer Interaction*, *26*(5), 537–561. Retrieved from https://doi.org/10.1080/10447311003720001 doi: 10.1080/10447311003720001

Story, M. F., Mueller, J. L., & Mace, R. L. (1998). *The Universal Design File: Designing for People of All Ages and Abilities. Revised Edition.* Center for Universal Design, NC State University, Box 8613, Raleigh, NC 27695-8613 ($24). Tel: 800-647-6777 (Toll Free)' Web site: http://www.design.ncsu.edu. Retrieved from https://eric.ed.gov/?id=ED460554

Straumsheim, C. (2017, mar). Berkeley Will Delete Online Content. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/news/2017/03/06/u-california-berkeley-delete-publicly-available-educational-content

Treviranus, J. (2014). *The Value of the Statistically Insignificant.* Educause Review.

Treviranus, J. (2016). Life-long Learning on the Inclusive Web. In *Proceedings of the 13th web for all conference* (pp. 1:1—-1:8). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2899475.2899476 doi: 10.1145/2899475.2899476

Treviranus, J. (2019). The Value of Being Different. In *Proceedings of the 16th web for all 2019 personalization - personalizing the web* (pp. 1:1—-1:7). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3315002.3332429 doi: 10.1145/3315002.3332429

Trewin, S. (2018a, nov). AI Fairness for People with Disabilities: Point of View. Retrieved from http://arxiv.org/abs/1811.10670

Trewin, S. (2018b, nov). Will AI Methods Treat People with Disabilities Fairly? *Medium, MIT-IBM Watson AI Lab*. Retrieved from medium.com/@MITIBMLab/will-ai-methods-treat-people-with-disabilities-fairly-7626b38f9cb5

Trewin, S., Morris, M. R., Azenkot, S., Branham, S., Bleuel, N., Jenkins, P., . . . Lasecki, W. (2019). *AI Fairness for People with Disabilities.* Retrieved from https://assets19.sigaccess.org/ai_fairness_workshop_program.html

Tscheligi, M., Krischkowsky, A., Neureiter, K., Inkpen, K., Muller, M., & Stevens, G. (2014). Potentials of the "Unexpected": Technology Appropriation Practices and Communication Needs. In *Proceedings of the 18th international conference on supporting group work* (pp. 313–316). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2660398.2660427 doi: 10.1145/2660398.2660427

UN General Assembly. (2007). *Convention on the Rights of Persons with Disabilities : resolution / adopted by the General Assembly, 24 January 2007, A/RES/61/106.* Retrieved from https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/convention-on-the-rights-of-persons-with-disabilities-2.html

US Department of Justice Civil Rights Division. (2006). *COMMONLY ASKED QUESTIONS ABOUT THE AMERICANS WITH DISABILITIES ACT AND LAW ENFORCEMENT.* Retrieved 2019-

08-30, from https://www.ada.gov/q{&}a{_}law.htm

von Schrader, S., Malzer, V., & Bruyère, S. (2014, dec). Perspectives on Disability Disclosure: The Importance of Employer Practices and Workplace Climate. *Employee Responsibilities and Rights Journal*, *26*(4), 237–255. Retrieved from https://doi.org/10.1007/s10672-013-9227-9 doi: 10.1007/s10672-013-9227-9

Wastfelt, M., Fadeel, B., & Henter, J.-I. (2006, jul). A journey of hope: lessons learned from studies on rare diseases and orphan drugs. *Journal of Internal Medicine*, *260*(1), 1–10. Retrieved from http://doi.wiley.com/10.1111/j.1365-2796.2006.01666.x doi: 10.1111/j.1365-2796.2006.01666.x

Wilde, D., & Marti, P. (2018). Exploring Aesthetic Enhancement of Wearable Technologies for Deaf Women. In *Proceedings of the 2018 designing interactive systems conference* (pp. 201–213). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3196709.3196777 doi: 10.1145/3196709.3196777

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How Algorithms Discriminate Based on Data they Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy*, *8*, 78. Retrieved from https://www.jstor.org/stable/10.5325/jinfopoli.8.2018.0078 doi: 10.5325/jinfopoli.8.2018.0078

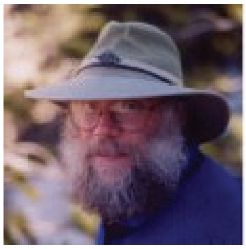
Wobbrock, J. O., Kane, S. K., Gajos, K. Z., Harada, S., & Froehlich, J. (2011, apr). Ability-Based Design. *ACM Transactions on Accessible Computing*, *3*(3), 1–27. Retrieved from http://portal.acm.org/citation.cfm?doid=1952383.1952384 doi: 10.1145/1952383.1952384

Zhang, S., Zhang, C., & Yang, Q. (2003, may). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5-6), 375–381. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/713827180 doi: 10.1080/713827180

Zolyomi, A., Ross, A. S., Bhattacharya, A., Milne, L., & Munson, S. A. (2018). Values, Identity, and Social Translucence: Neurodiverse Student Teams in Higher Education. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 499:1—-499:13). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3173574.3174073 doi: 10.1145/3173574.3174073

Shari Trewin is a technical leader in IBM's Accessibility Leadership Team, an ACM Distinguished Scientist, and Chair of ACM SIGACCESS. She has a Ph.D. in Computer Science and Artificial Intelligence from the University of Edinburgh, with 17 patents and over 60 publications in technologies to remove barriers for people experiencing disabilities.

Sara Basson works at Google as "Accessibility Evangelist," with the goal of making the experience of Googlers more accessible and usable. She previously worked at IBM Research on Speech Technology, Accessibility, and Education Transformation. Sara holds a Ph.D. in Speech, Hearing, and Language Sciences from The Graduate Center of CUNY.



Michael Muller is an ACM Distinguished Scientist, and a member of the SIGCHI Academy with expertise in participatory design and participatory analysis. In the AI Interactions group of IBM Research AI, he focuses on the human aspects of data science; ethics and values in applications of AI to human issues.

Stacy Branham is an Assistant Professor of Informatics at the University of California, Irvine. Her research explores the role of technology in social integration for people with disabilities, with active lines of inquiry in the domains of personal mobility and parenting with disability.

Jutta Treviranus is the director and founder of the Inclusive Design Research Centre (IDRC), established in 1993. The IDRC leads many international research networks to proactively model more inclusive socio-technical practices, including in the area of data science (e.g., Project We Count). Dr. Treviranus has co-designed numerous policies related to digital inclusion.

Dan Gruen is a Cognitive Scientist in IBM Research's Center for Computational Health focusing on AI explainability, clinical reasoning, and complex medical cases. Dan has a PhD in Cognitive Science from UCSD and holds over 50 US and international patents in User Experience, Visualization, Social Software, and related areas.

Daniel Hebert has his computer science degree from Rensselaer Polytechnic Institute and has been with IBM since 2006 as a QA, engineer, and developer. He enjoys working on smart home and accessibility projects, and is a community advocate for persons with disabilities and involved with the local independent living movement.

Natalia Lyckowski is an Application Developer with IBM Global Financing. She is also a passionate volunteer educator on inclusion and diversity awareness. Most proudly, she is the mother of a young autistic man who has taught her how to think out of the box and remain ever flexible.

Erich Manser is an IBM Accessibility evangelist exploring new approaches to accessible technology, new applications of emerging technologies, and contributing to accessibility standards. Beyond the workplace, Erich is an active mentor, disability rights advocate, and fundraiser. He is also a nationally recognized blind athlete in marathon and triathlon racing.

The Intersection of Ethics and AI

Annie Zhou (High Technology High School; anniezhou35@gmail.com)

DOI: 10.1145/3362077.3362087

Introduction

Artificial intelligence is a rapidly advancing field with the potential to revolutionize health care, transportation, and national security. Although the technology has been ubiquitous in every day society for a while, the advent of self-driving cars and smart home devices have propelled a discussion of the associated ethical risks and responsibilities. Since the usage of AI can have significant impacts on people, it is essential to establish a set of ethical values to follow when developing and deploying AI.

Responsibility

The companies researching, developing, and deploying artificial intelligence must be held accountable for their products. With the future of artificial intelligence as vastly uncharted territory, companies must be mindful of their corporate social responsibility: namely, the impact of their products on society in various areas (Polli, 2017). AI research may be rapid, but companies must hold back before quickly releasing new products to ensure they are following ethical behavior. Several tech giants have acknowledged the significance of their role in the future of AI; Microsoft, IBM, and Google, among others, are building standard ethics processes. Smaller companies have to also begin assembling a code of AI ethics that sets standards and precautions to deal with various issues in a consistent manner (Hao, 2018).

In order to analyze the ethics of the technology, corporations need to shift infrastructure to seamlessly include ethics in AI decisions. For one, they need to hire ethicists who can advise the programmers. Corporations also need to create ethics training programs to teach programmers about the ethical concerns of AI, so the ones work directly with the technology understand how they're affecting society.

Corporations will also need to set rules in the case their AI technology inflicts harm. The farther the technology goes from the algorithms, the more unclear it is on the companies' liability (Baker, 2004). For now, property damages and harm will have to be examined on a case to case basis. The company will be held liable in instances their product fails due to an oversight in production, just as if products were built with a poor design and hurt the user, the company is responsible.

In March 2018, an experimental Uber self-driving car struck a pedestrian, after a coded system for emergency stops was disabled (Ford, 2018). In the case this occurs when self-driving cars are implemented nationwide, possible lawsuits can follow the model of product liability precedents. Liability can be divided into strict liability, negligence, manufacturing defects, and design defects (Villasenor, 2018). In the case that a defective product is released, then the company is liable for any resulting damage. If the company failed to test the product in a multitude of possible circumstances, such as testing braking systems only on dry services, and the product subsequently fails on wet roads, the company is liable for crashes caused by wet roads. If a manufacturing issue results in damage, then it is the fault of the intermediary manufacturer. If the product is fundamentally designed in a way that incites harm, the company is responsible for addressing this issue (Lea, 2018).

The government also ought to play a role in regulating an ethical implementation of AI systems. As AI makes consequential decisions about people, the government will need to develop and provide public policy to ensure AI is for the public good. Several measures have been taken, such as creating a subcommittee on the National Science and Technology Council for Machine Learning and Artificial Intelligence. The government ought to encourage datasets that are representative and fair, by releasing government data sets to aid AI research and creating open data standards. Government programs that will be affected by AI, such as the Department of Transportation, should work with researchers. A government committee can also regularly monitor AI research progress and meet with industry.

Transparency & Database Bias

Algorithms will need to be developed in a transparent manner. Consumers will need to establish their trust in AI systems, beyond simply the accuracy of their desired function. The code itself may seem long and arduous to understand for the general public. The IBM team recently proposed four fundamental pillars for trusted AI: robustness, explainability, lineage, and fairness. (10153623502710135, 2018). AI models need to strike a balance between explainability on how they arrive to specific decisions, and accuracy. Meaningful explanations about AI models reduce uncertainty. AI systems must be safe in the sense that it follows laws, societal norms, and regulations associated with safe behavior. They must also be secure from adversarial attacks. Malicious actors can alter databases to influence the behavior of AI models, which can lead to inaccurate or biased results. Companies must account for possible attacks and test their models against adversarial attacks. Understanding the lineage of an AI system, namely its history and past configurations, is important for trusting in AI. One of the most major concerns is the fairness of AI systems.

As AI systems rely solely on learning from a database, if the database is incomplete or biased, the system can unintentionally perpetuate bias to a widespread scale. Unrepresentative data can latch onto subtle racist and sexist patterns. For example, risk assessments provided by computer programs were used to predict the likelihood of a criminal committing a future crime. A ProPublica investigation discovered that not only was the risk score unreliable for forecasting repeated crimes, the formula was twice as likely to flag black defendants as future criminals, and white defendants were mislabeled as low risk (ACLU, n.d.). This disparity stemmed from the existing bias in historical records of arrests, from which the data was used to teach this algorithm. It is highly dangerous for criminal justice algorithms to be based on bias, as that violates the justice system's fundamental goal. Biased algorithms could perpetuate a vicious cycle where more African Americans are unfairly incarcerated.

Bias was also evident in Amazon's recent system for automating the recruitment process. Since the tech industry was male-dominated, most of the resumes fed into the machine learning model were for men. The resulting system favored men over women. If this recruiting system were deployed, it would have further alienated women from the tech sector.

AI Interaction with Humans: Customer Engagement

As new AI-technologies enter the market, the prevalence of AI only grows. Although a common fear is that AI will eventually replace the human race, AI is expected to be complementary to humans. One particular field where AI is being deployed in is customer engagement. 38% of enterprises are planning to experiment with AI powered customer service, according to a Gartner survey, AI has the potential to give customers the right information they need, while freeing up time for customer service representatives to handle more complicated issues.

In 1950, Alan Turing introduced the Turing test in which judges blind conversations with a chatbot and human participants. Based on the responses, the judges had to identify which conversation was with a chatbot and which was with a human. The test was constantly analyzing what it meant to be human versus a machine interpretation of humanity. The results of the annual test revealed that humans had several distinctions in their speech, such as timing, fluidity, and vocal tics such as "uh" and "mhm". Google recently unveiled Google Duplex, an AI that can make a phone call that sounds eerily similar to one a human would make, with natural pauses, varied questions, and other idiosyncrasies indicative to human speech (Vincent, 2018). This development raised concern over whether Google Duplex had an obligation to tell the people it called that they were talking to a machine. Google Duplex also raised the ethical question of making a robot voice that could be easily mistaken for a human's. Though Google has promised regulations to ensure that their usage of the technology is solely for making acts like restaurant reservations easier, this technology is so dangerous. Robots that can pose at humans could pave the way for a new wave of scam calls and hoaxes.

Regulations must be made for automated AI conversations. Before starting a call, the AI should warn the person it's calling that an AI is speaking. AI conversations should also be limited to following simple directives such as ordering gifts or making a reservation. AI should not be pushed to be able to hold full conversations indistinguishable to ones humans hold, as such technology could be exploited to trick people into thinking they're speaking a human. Although some people will automatically hang up once the AI announces its identity, companies can works towards calming the general public's fear over the technology by being completely transparent about what their algorithm intends to do.

AI Interaction with Humans: Self Driving Cars

Vehicle automation is close to becoming a reality on the roads, with companies like Mercedes-Benz, Tesla, Toyota, and Uber working on driverless car technology. When entrusting the decisions on the road to a computer program, several ethical concerns can arise. When a machine makes decisions, it has to follow a preprogrammed standard of ethics. The most popular dilemma is what decision to make if a crash were imminent. If a human driver slams on a brake to avoid hitting a pedestrian crossing a road illegally, they may be putting the passengers in danger. However, as it is ethically gray whether it is better to save the pedestrian or the passengers in side the car, it is controversial what decision self-driving cars should make. As different people have different values, being forced to choose one universal moral code for self-driving cars to follow is difficult. In a split second decision, a human would typically lean towards preserving his own protection. If the decision can be pre-programmed and pre-ruminated, results may vary. A paper published by MIT details an online quiz called the Moral Machine, where it asked users a series of ethical choices regarding hypothetical car crashes, in a manner similar to the classic trolley problem. Certain trends were apparent, such as sparing humans over animals and generally favoring more lives than fewer, but people were split over the crucial decision whether to save someone younger or older or a pedestrian versus a passenger. Germany's Ethics Commission on Automated Driving published propositions that other countries can follow as a guide. These included affirmations that human life takes top priority and that any distinction in personal features in possible victims of an unavoidable accident situations cannot be considered in calculating what decision to make. Generally, it would be optimal for cars to be programmed with automatic risk calculations to weigh the the morality of certain decisions. For example, if a person were crossing a street illegally, they are doing so with the knowledge that this is illegal, so his life would be prioritized underneath the passenger's life.

These moral quandaries are extreme, and are highly unlikely to occur in practice. There are more common ethical problems that self-driving cars will inevitably run into. Driving well not only involves driving safely, but also taking smart risks. For example, a driver wanting to merge onto a highway may either choose a safe speed, or a faster one to reduce overall travel time. When following the flow of traffic, cars often go close or over the speed limit; slow, cautious cars are not helpful for traffic flow. Autonomous vehicles will need to be equipped to weigh decisions based on the risk and the reward, striking a balance between safety and efficiency. Furthermore, the behavior of cars will influence future traffic patterns. In anticipation of a future where all transportation is automated, subtle driving behavior patterns programmed into the algorithms will become the norm. Thus, every small decision has to be considered carefully regarding its impact in the traffic scheme (Bogle, 2018).

Overall, if self-driving cars expected to become the norm, they must be built in a transparent manner. In order for consumers to trust this innovative technology, they must be able to know everything about how the algorithms are made. Collaborations among social scientists and ethicists with programmers must be fostered, so that these ethical issues are addressed.

Consequences on Economy

Autonomous vehicles have the potential to be the future of all transportation. Switching to self-driving cars would significantly impact the economy. Out of the most common professions per state, truck driving is the most common. In 2014 around 1.6 million Americans were employed as truck drivers (Service, 2017). Around a million people are employed as taxi drivers, Uber drivers, school bus drivers, and transit drivers. If the need for drivers is eliminated, then all these people are out of a job. It extends beyond the drivers. Others jobs can be affected, such as gas station attendants, rental car agencies, street meter maids, and repair shops. Up to 4 million people are at risk if full automation is achieved. Not all these people will be unemployed; with the natural expansion of the economy, they will inevitably find jobs in the new industries created from automation. However, companies unveiling this technology ought to take several precautions. There are several possible ways to smoothly transition the economy in preparation of autonomous vehicles. For one, they can develop a new profession in the form of remote vehicle operators. Companies will likely establish command centers to monitor the autonomous vehicles. Another idea is a “passenger economy”, with goods and services being provided to people during their ride in an autonomous taxi. The government can also collect money from the self-driving car industry, and use it to help re-position the people affected into new jobs. This money can help tide them over until they find a job, or it could be used to create jobs for people in the form of building civil works projects.

Cultural Differences Regarding AI

Artificial intelligence is a technology being explored globally. People from different cultures have different opinions on how to deal with it. This depends on the different opinions of “humanity”. For examples, follower of Shinto, the official national religion of Japan, believe that humans are just the same as rocks and animals, as everything is part of Nature. On the other hand, Western philosophy prides in the individuality and uniqueness of humans, and draws a distinct line between what is human and what is not. The Western fear of robots largely stems from the fear of pushback from what humans have dehumanized. As each culture has different perspectives of artificial intelligence, it is important to coordinate these different opinions. It will take global cooperation to set ethical guidelines for AI technologies (Ito, 2018). Countries should engage in discourse on relevant artificial intelligence issues, introducing the latest developments from their countries and promoting collaboration of scientists from different countries. Although it is impractical to try and change the perspectives of people from different cultures, it is important to understand why other people view artificial intelligence a certain way, in order to create appropriate legislature.

Teaching AI Ethics

In order to ensure that ethics are considered when developing artificial intelligence, the general public needs to mindful of the importance of ethics in this technology. Cooperation between computer scientists and ethicists is essential. These two groups of people can collaborate to come up with curriculum to teach people the intersection of ethics and AI. This curriculum can be taught through courses at universities, in order to teach the upcoming generation of leaders in this field.

Programs can also be developed to start teaching the intersection of artificial intelligence and ethics at a young age. Summer programs like AI4ALL reach out to high schoolers, introducing them to artificial intelligence in the context of the various ethical issues involved. These programs can be further implemented through online courses, reaching a greater audience. As artificial intelligence is especially pertinent to the government, the government can fund programs to teach the ethics of AI. Social experiments such as the Moral Machine are also effective in alerting people to the possible ethical dilemmas that come along with artificial intelligence (Maxmen, 2018).

Overview

Artificial intelligence has proven to be a relevant technology that will exponential grow in future years. Although the technology has great potential, it raises many ethical issues regarding its deployment. It is important that companies developing the technology are mindful of their ethical responsibility in developing the technology, and that the government creates appropriate regulation. Algorithms will need to be transparent and based upon unbiased, complete data. Artificial intelligence must be for the common good, not perpetuating current biases in society. Furthermore, the general public will need to trust artificial intelligence's reliability, security, and safety. As artificial intelligence increasingly matches human behaviors rules must be put in place to avoid people misusing its human mimicking ability. Ethical concerns are especially apparent regarding self-driving cars, which will need to follow an established set of rules. The consequences of AI on the economy will need to be considered. Global cooperation is necessary to successfully implement AI. Overall, the intersection of AI and ethics is crucial to its success, and companies and the government should aim to educate the public about the implications.

References

10153623502710135. (2018, October 08). Towards AI Transparency: Four Pillars Required to Build Trust in Artificial Intelligence Systems. Retrieved from https://towardsdatascience.com/towards-ai-transparency-four-pillars-required-to-build-trust-in-artificial-intelligence-systems-d1c45a1bdd59

ACLU. (n.d.). With AI and Criminal Justice, the Devil Is in the Data. Retrieved from https://www.aclu.org/issues/privacy-technology/surveillance-technologies/ai-and-criminal-justice-devil-data

Ark, T. V. (2018, September 13). Let's Talk About AI Ethics; We're On A Deadline. Retrieved from https://www.forbes.com/sites/tomvanderark/2018/09/13/ethics-on-a-deadline

Artificial Intelligence raises ethical, policy challenges – UN expert — UN News. (n.d.). Retrieved from https://news.un.org/en/story/2018/11/1025951

Baker, M. (2004, February 06). Definitions of corporate social responsibility - What is CSR? Retrieved from http://mallenbaker.net/article/clear-reflection/definitions-of-corporate-social-responsibility-what-is-csr

Bogle, A. (2018, March 21). 5 big ethical questions about driverless cars we still need answered. Retrieved from https://www.abc.net.au/news/science/2018-03-21/self-driving-autonomous-cars-five-ethical-questions/9567986

Bundy, A. (2016). Preparing for the future of Artificial Intelligence. Ai & Society,32(2), 285-287. doi:10.1007/s00146-016-0685-0

Burton, E., Goldsmith, J., & Mattei, N. (2018, August 01). How to Teach Computer Ethics through Science Fiction. Retrieved from https://cacm.acm.org/magazines/2018/8/229765-how-to-teach-computer-ethics-through-science-fiction/fulltext

Ethics Commission on Automated Driving presents report. (n.d.). Retrieved from https://www.bmvi.de/SharedDocs/EN/Press Release/2017/084-ethic-commission-report-automated-driving.html

Ford, M. (2018, March 20). A Self-Driving Uber Killed a Woman. Whose Fault Is It? Retrieved from https://newrepublic.com/article/147553/self-driving-uber-killed-woman-whose-fault-it

Hao, K. (2018, October 21). Establishing an AI code of ethics will be harder than people think. Retrieved from https://www.technologyreview.com/s/612318/establishing-an-ai-code-of-ethics-will-be-harder-than-people-think/

Ito, J. (2018, July 31). Why Westerners Fear Robots and the Japanese Do Not. Retrieved from https://www.wired.com/story/ideas-joi-ito-robot-overlords/

Lea, G. (2018, December 06). Who's to blame when artificial intelligence systems go wrong? Retrieved from https://theconversation.com/whos-to-blame-when-artificial-intelligence-systems-go-wrong-45771

Lee, J. (2015, June 19). Self Driving Cars Endanger Millions of American Jobs (And That's Okay). Retrieved from https://www.makeuseof.com/tag/self-driving-cars-endanger-millions-american-jobs-thats-okay/

Maxmen, A. (2018, October 24). Self-driving car dilemmas reveal that moral choices are not universal. Retrieved from https://www.nature.com/articles/d41586-018-07135-0

Pakzad, R. (2018, February 01). Artificial Intelligence and Corporate Social Responsibility. Retrieved from https://medium.com/humane-ai/artificial-intelligence-and-business-social-responsibility-69d6299b4d9d

Polli, F. (2017, November 09). AI And Corporate Responsibility: Not Just For The Tech Giants. Retrieved from https://www.forbes.com/sites/fridapolli/2017/11/08/ai-and-corporate-responsibility-not-just-for-the-tech-giants/790804f95d4b

Responsibility in Technological Civilization:In Search of the Responsible Subject. (2013, October 25). Retrieved from http://www.iwm.at/publications/5-junior-visiting-fellows-conferences/vol-xxx/anastasia-platonova-responsibility-in-technological-civilization/

Service, N. Y. (2017, December 15). Major job losses - and gains - loom with autonomous vehicles. Retrieved from https://www.boston.com/cars/car-news/2017/12/15/major-job-losses-and-gains-loom-with-autonomous-vehicles

Social Responsibility and Ethics. (n.d.). Retrieved from https://www.pachamama.org/social-justice/social-responsibility-and-ethics

The Ethical Challenges Self-Driving Cars Will Face Every Day. (2018, March 27). Retrieved from https://www.smithsonianmag.com/innovation/ethical-challenges-self-driving-cars-will-face-every-day-180968596/

Villasenor, J. (2018, May 09). Products Liability and Driverless Cars: Issues and Guiding Principles for Legislation. Retrieved from https://www.brookings.edu/research/products-liability-and-driverless-cars-issues-and-guiding-principles-for-legislation/

Vincent, J. (2018, May 09). Google's AI sounds like a human on the phone - should we be worried? Retrieved from https://www.theverge.com/2018/5/9/17334658/google-ai-phone-call-assistant-duplex-ethical-social-implications

Walker, J. (n.d.). The Self-Driving Car Timeline – Predictions from the Top 11 Global Automakers — Emerj - Artificial Intelligence Research and Insight. Retrieved from https://emerj.com/ai-adoption-timelines/self-driving-car-timeline-themselves-top-11-automakers/

Who should be blamed when the machine makes the mistake? (2017, June 07). Retrieved from https://www.healthcareitnews.com/cloud-decision-center/who-should-be-blamed-when-machine-makes-mistake

Annie Zhou Annie is a senior at a rigorous pre-engineering vocational high school interested in computer science, design, and finance. She serves as her senior class president and finance club president. She is fascinated by the interdisciplinary potential of artificial intelligence and hopes to work in the ever-expanding AI field in the future.

Artificial Intelligence: The Societal Responsibility to Inform, Educate, and Regulate

Alexander D. Hilton (Southern New Hampshire University; alexander.hilton@snhu.edu)

DOI: 10.1145/3362077.3362088

Introduction

Artificial Intelligence (AI) is a rapidly growing field; one that is mysterious to the general public. The mention of the word AI fills the imaginations of many with thoughts of talking robots, jobs being replaced, and possibly even the destruction of mankind. Perhaps imaginations are running wild due to, perhaps driven by the loose definition of AI as systems able to perform tasks that normally require human intelligence that allows Hollywood to take some creative license. The experts in the field tend to work directly with AI and often for large companies, allowing for the imagination and news headlines to be where the public gets their information. Many wonder if this new technology is going to be an overall benefit to society or if it will bring unmitigated disaster. When the imagination runs wild, instead of understanding, news stories can perpetuate concerns and anxieties rather than hope and optimism.

Andrew Ng once said "Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI will transform in the next several years." It may well be the next electricity in the way it will revolutionize and change both industry and even our daily lives. From applications that identify faces in photos on social media to deep learning models meant to help discover new cancer treatments, the potential of AI can impact nearly every person's life and may already have to some degree. Most of us are already interacting with AI in some form just when we use Google or Facebook, often without our knowing. You may have interacted with one today without even realizing it. When faced with technology that has a large reach and broad scope, it is imperative to consider how the lives of millions could be impacted. We are already seeing individuals stepping up in the political realm with ideas of how to address safety and privacy concerns associated with AI. Andrew Yang, who announced his bid for the 2020 Democratic Presidential nomination, is running on a platform that stated 'the robots are coming' saying that AI will change virtually every part of the global economy. He could be right.

It is certain that AI will begin to play a more prominent role in our daily lives as the field develops and new uses are found for such systems. The change AI can influence our lives in positive ways but could put lives at risk if irresponsibly used. Unfortunately working through governmental bureaucracy and regulation can take time, yet AI continues to advance rapidly. As both individuals and corporations take huge strides forward some governments have begun to realize the impact of easily accessible data that feeds into AI. The European Union implemented the General Data Protection Regulation in 2018 as an effort to inform and gain consent by people to ensure their data cannot be exploited.

The General Data Protection Regulation (GDPR) was an attempt to begin to change how EU citizens interact with their data and offer some protections. AI is thriving in this era of big data, however, unlike data, people will be interacting directly with AI in new and intriguing ways. Yet these interactions will not just be predicated on the quality and behavior of the AI, but also the accessibility and knowledge of the human involved. Interactions are a two-way street and while we need to look at the regulations put on AI, perhaps the most important side of the equation is how society views and interacts with AI. Discussing these requirements going forward in this coming age of AI, which will likely cause a renaissance for society, must consider the humans involved and the technology simultaneously. The quality of the interaction that AI and humans will have will stem from knowledge, access, and open-mindedness.

Development of these interactions must come from establishing knowledge, establishing trust, and then encouraging responsible use. Allowing for this sort of positive interaction, however, takes more than just governmental oversight, but also investment and a commitment from society. Encouraging these interactions should start with the hardest factor to change. AI can be reprogrammed or retrained, humans are a bit more difficult to change. It is critical to encourage and invest in the human element first before discussing the requirements put on AI.

The Human Factor

Up until recently, I was someone who was about as versed in AI as the average American. Most of my knowledge came from Hollywood movies like “I, Robot”, “The Matrix”, and “Terminator” where AI is depicted as programmed systems coders wrote logic for, which ended up leading to some devastating results. Then I saw the incidents making national headlines of AI crashing cars, chat bots that respond with racist remarks, or smart TVs and speakers listening into conversations and I began to worry.

Most people chalk these problems up to human error, logical issues we could not foresee making it hard to code. Sure, a team of programmers could correct these mistakes, but what about the human error caused by that correction. Humans are fallible so would the hard-coded logic I viewed AI as be the same? Sure, I was optimistic about the prospects of AI, but what if these sci-fi stories and futurists were more foretelling than just stories? Could a rogue AI change everything about our lives for the worse? Human error will not go away, even the best program can have a bug, AI did not seem it would be different. It made me anxious, and I empathize with people who still feel that way.

Studies have shown many Americans too view artificial intelligence in a similar light. While people tend to be cautiously optimistic about how AI can positively impact their lives, they are also very anxious about the prospect of AI. Many people worry about it changing the industries around them, perhaps taking the jobs of their friends and family. There is concern that AI cannot handle lives as safely as a human, an anxiety which is increased when we hear of car accidents involving AI and the loss of life. These anxieties are real and as AI becomes increasingly prevalent these breaking news stories won’t just happen a few times a year, but could occur daily just due to how available AI is becoming. Much of these anxieties revolve particularly around jobs and losing jobs to robots and AI automation (Gallup, 2018).

Recently, however, I’ve been able to put my own anxieties regarding artificial intelligence aside. The reason for this was simple, although it was also a bit of a journey. As a data analysis student, I became curious about machine learning and deep learning in their roles in analysis which led me to further my education on the matter. The resources I found came from people like Andrew Ng, Siraj Raval, and Cognitive Class from IBM. Many of these people and classes came at no cost which made it very easy to dive into without having to worry about the financial commitment. It blew me away to learn that most AI we discuss today is not hard coded logic, but rather mathematical operations performed on massive amounts of data. These algorithms, while still susceptible to human error and issues in the data, are learning models that can improve as our data processing abilities become better. Likewise, with many of these neural networks being open source packages like Tensorflow and PyTorch, a massive community of engaged data scientists and programmers can improve how neural networks and AI is designed, in many ways democratizing the process. There isn’t just some small group of mad scientists trying to make a humanoid robot, but rather a substantial community that is engaged in trying to make the most of AI.

Neural networks as well as much of modern artificial intelligence can learn and improve far easier than I imagined and the community behind much of AI wants to encourage society to take that next step into the future with an open mind. AI certainly has the capacity for replacing jobs or crashing cars, but these AI can be improved to facilitate our lives offering new opportunities even for the very jobs that they are said to replace. The ease of accessibility to knowledge and the community that discusses and works on building artificial intelligence and neural networks shows that people want to improve these systems as well as improve the lives of others. After learning about the easily accessible and free learning options pertaining to AI, my anxiety turned more optimistic as worry turned into hope.

Knowledge is power, especially when that knowledge is easily accessible. As fear abates the full capabilities of AI can be revealed. When new technologies for instance are introduced there are always anxieties regarding the new unknowns that have been introduced. When electricity was first taking off, there was often fear related to what it could do. Some claimed it could even destroy the concept of day and night, which some thought could significantly impact our culture, or even our health. While some of these fears could be stoked by incidents like electrical fires, allowing people to become more educated about how electricity works curbs these fears. People can learn how to protect themselves from starting electrical fires with some education growing up with the technology, as well as being told not to shove metal object in the socket. With AI, the principles are much the same, people need to be taught how to adequately protect themselves from harm rather than just give into fear and myth put out by movies or sensationalized news headlines.

However, recent incidents have shown how society reacts to irresponsible use of technology, especially regarding big data which is closely tied to AI. Cambridge Analytica was infamous for using data and machine learning to try to change and influence elections, most notably in Kenya where they used their insights to re-brand a certain political party twice (Lang'At, 2018). People who were unaware that their data was being used were easy to target. These individuals could be advertised to or presented with information that could have changed their political opinions. When the story broke about how Cambridge Analytica had gathered so much personal information on people globally, the US Senate had hearings and the EU introduced the GDPR in response as well.

The reaction to the news stories regarding the Cambridge Analytica scandal was not unexpected, but it did reveal a lot of societies lack of knowledge when it came to these issues which the GDPR sought to address. While this was a regulation, the idea of informed consent regarding using data is based on establishing a degree of trust and interactions with data. This concept plays into how AI functions as AI requires data to learn. While these regulations could hinder innovation, it is a key step into establishing trust in these interactions as well as ensuring people be informed about the systems they are interacting with. If humans interacting with devices that contain AI are unaware that these devices are gathering and processing their data, even in a benign way, if (and when) that data gets leaked then there is a huge breach of trust. When humans are caught off guard and made distrustful of these devices, single-event learning can take hold which can turn off people to AI entirely. People do not regain trust quickly so avoiding this sort of situation is paramount.

Nevertheless, simply informing people that they are interacting with AI is only a first step. Being informed is the start to education but going a bit further can put those anxieties to rest while opening doors for new opportunities. Presenting educational opportunities for people to learn how AI works and how they learn and improve is a start. When people are introduced through free resources by some engaged minds and companies in the field, it lets them explore and learn about these systems in an empirical manner. Education has always been a good tool to helping with anxieties that involve natural and artificial phenomena, people are not terrified Zeus will strike them down with a bolt of lightning now that we have a general understanding about weather patterns for instance. Discussing AI should be treated similarly especially as it becomes increasingly more prominent in our day-to-day lives. Now this isn't to say that everyone needs to get a Master's in Computer Science (nice as that would be), but that people have a similar understanding of AI as they do about electricity for instance.

Spreading public general awareness of AI through free, or at least very cheap, educational programs allows for an open approach. While it would be nice to see these groups have a larger reach, they are forging a path in the right direction. Keeping the knowledge out there and accessible to the public helps ensure that people will start to learn their own best practices as they watch AI grow. As people start becoming more aware of the AI in their lives and have an easy ability to learn about those systems, then gradually society can adopt AI into their lives responsibly just like we have done with electricity.

Furthering an open approach to help spread public awareness and knowledge should be considered a societal investment. AI has a significant amount of potential, however, if people have anxiety and fear while simultaneously being uninformed can allow for society to create bad policy and bad practice that stifles innovation and growth. While groups are gradually informing the public, a proper investigation into a public awareness campaign and increasing public education on the matter should be considered by industrialized nations. Likewise, investments in education and higher education are a must going forward, not just for people to understand how AI works, but because of its potential to change the job market as well.

Fostering education in AI can come from non-profits, companies, or even governments. Groups like the ACM SIGAI sending out newsletters and offer webinars to members is just an easy way to keep people informed of what is going on in the field and what to be aware of. Other groups do this as well such as Andrew Ng's deep learning.ai and Siraj Raval's School of AI. Even IBM's Cognitive Class offers free classes for people to learn programming with machine learning and deep learning. These free and inexpensive courses do not preclude anyone from learning about AI. This approach makes learning about AI easier for those who are interested, these groups are all easy to approach and often have a fair bit of free content to help people become educated and even versed in the various aspects of AI, regardless of income. As well, people who post on Github can share their knowledge and expand the reach of understanding. As stated before even the open source packages like Tensorflow open many opportunities by encouraging people to work with deep learning, while Google still profits through their cloud and selling their tensor processing units, the overall accessibly Tensorflow has offered allowed many people to experiment with and learn about deep learning hands-on.

Expanding the accessibly to and education of AI should be a focus; some nations are working to that end. China is attempting to get a one trillion-dollar industry developed. The EU is putting billions of dollars of investment into AI to try to catch up with the US and Asian countries. MIT is in the midst of building a billion-dollar AI college to make AI part of every graduate's education. All this investment will hopefully allow more members of society to be informed about and able to adequate use and interact with artificial intelligence. Likewise, this increases the availability of experts and enthusiasts that can then go out to develop businesses, large and small.

Taking steps to ensure that AI businesses are invested should be a priority. Ensuring that these new technologies are subsidized to be easily accessible by the wider community rather than just being proprietary is important as well as AI then becomes more like a utility rather than a luxury. The accessibility is vital to ensuring that people can start to become educated. Finally, investing in education whether it be through public education initiatives or just through communities forming together to educate people should be encouraged by society in both the form of private investment and governmental assistance.

Though, with all these enthusiasts and business there is a chance mistakes will be made such as the events Cambridge Analytica helped spark. Especially when AI can produce "deep fakes" or other forms of scandals, it should be expected that we will see some horror stories in the coming decades. With an educated populace, the risk of panic and fear can be mitigated substantially and trust in AI will not be broken entirely. Likewise, the appropriate regulations can be discussed early without concern of over regulating and stifling the development of better artificial intelligence.

Regulation and Requirements on AI

For any decisions with AI interacting with humans, the best solution is allowing both society and AI to treat those interactions as learning experiences. Ensuring access to education is a first step here along with ensuring accessibility to AI. This accessibility allows AI to be treated more like a utility which can help minimize damage when accidents occur and the fear generated by those mistakes.

So long as these inevitable errors are small and do not put lives in serious jeopardy, the hope should be that the AI community and the greater global society can inspect and adapt as issues crop up with AI. However, in order to properly carry out this inspection society must be knowledgeable of what they are interacting with.

As we saw with the Cambridge Analytica case and the GDPR that the EU put in place in response, similar mechanisms should be considered for AI. When the GDPR came into effect, two out of three people felt more comfortable sharing their data (Association, 2018). This statistic alone is important for AI as people who feel more comfortable sharing their data might be more willing to hand that data over, so it can be used for training more complex models. In some ways, despite how the GDPR has been suggested to impact businesses, it may be necessary to the long-term trust in the data collection process, which is incredibly beneficial for AI.

In a similar way, as the GDPR helped build confidence in people's use of data, so to could be similar regulation to keep people informed when they are interacting with AI. There are no surprises that way which increases faith in these systems through informed consent. Perhaps a GDPR for AI is the next logical step for nations to consider. It increases society's faith in these technologies which may mean more data and information can be gathered through those interactions, thus leading to more information to improve the next generations of AI. This regulation could be as simple as just labeling something as containing AI or could go into more detail, such as if a neural network is gathering video data and whether that network is a convolutional neural network or if it is recurrent network which could help identify what the AI is trying to accomplish. Informing people of the process can let them know whether they are dealing with an object recognition program or perhaps one that monitors and identifies activities. Different people might feel better about interacting with one system versus the other, thus should be aware of what they are handling. So long as people are generally aware of what the artificial intelligence is doing, they can volunteer to still work with and use it rather than worry that the AI is doing something it is incapable of doing.

Trust is key to the long-term usefulness and viability of AI in general. Building and maintaining trust in these systems as mistakes are made will ensure society does not treat AI with concern and anxiety, but as a tool to be utilized to benefit and grow our lives. In the cases of human lives being put at risk as well, trying to ensure that informed consent is given at least creates an understanding of the potential risks as well. These stories become less terrifying as the onus still falls to the human working with the AI rather than people just blaming and worrying about the AI.

Yet society should not simply accept any company or individual putting lives needlessly at risk. It is an advantage to encourage small businesses and individuals to harness AI as the ramifications can be less than if larger companies are the ones to make these mistakes. We know this from examples as well. These mistakes won't be like the Equifax breach in scale, but they can still happen and even a simple car accident with an AI driver can make national headlines. We cannot necessarily expect an AI to be working at Bayes error rate, the lowest possible error rate, every single time. In these situations where lives can be immediately put at risk; the AI should be making errors at similar levels to humans before they are deployed, thus needless risk is not introduced. While this can limit innovation, there are potential work arounds, such as with self-driving cars having a human driver as well. This system decreases the chance of an accident, but mistakes can still happen as we saw in 2018 with a self-driving car crash in Arizona where a human driver was still involved. That incident shows that a higher standard must be set.

Where lives are directly involved AI should only be deployed if the error rate the AI makes is at the same level of humans or preforms better than a human. We can expect AI to make errors, but the goal should be to minimize that error rate as much as possible. The struggle is that these systems often learn through gathering massive amounts of data which means at some point they need to be tested in real-life scenarios. Under these circumstances, the AI must have that error rate verified before they enter these situations that could put lives at risk. Much in the same way we impose driver's licenses on people, so too should AI that could lead to loss of life. Once the human error rate is met, then the AI should be deployed. At this point it can gather data and learn about these real scenarios instead of just learn on simulated data. With any luck, the AI will surpass human error as it learns more in real world scenarios (Jalan, 2017). Perhaps even in the incidents that do led to tragedy, the AI will still improve past the human.

There are even a few potential considerations to have. A self-driving car cannot get drunk and could save lives in the end by preventing incidents like DUIs which can cause thousands of deaths every year (CDC,). Another area could be AI driven drones that, if properly programmed and trained, can be used to target enemy combatants and work to prevent civilian casualties. AI does not have the trouble with fatigue and if it misidentifies, drones powered by AI could take our troops out of harms way while simultaneously making less mistakes such as friendly fire. Along the same lines, long haul trucking or flights could be piloted or assisted by AI without the concern of fatigue or illness potentially imparting cognitive functions. Regulating for what we have never seen or given the opportunity to try though could shut down opportunities to improve the human condition and these AI technologies. The potential opportunity for AI could lead to more lives saved in the end, so long as we can learn the appropriate applications and limitations that is.

Imposing regulation where lives are at risk could be built around the very same concepts used to license human drivers and operators. This regulation would be the most intense as it means creating a system and licensing agency. Previous to the age of big data and the ability to create simulations, this may have been impossible. However, now we can, with a fair bit of accuracy, compare how an AI stacks up to a human in a variety of situations. When the error rates are roughly the same in simulated environments, the company or individual can request for an operating permit to deploy these AI in a real context. While this is massive government oversight, it ensures lives are not put at wanton risk just for the sake of progress or a quick buck.

With these two requirements put on AI, one that requires informed consent for these interactions and another to ensure that where lives are at stake the risks are mitigated, maybe society can start to build trust with these new systems and as more complex AI enter our lives. Perhaps society would even welcome more human-like robots into their lives as these anxieties are curbed through knowledge, consent, and trust.

Conclusion

The key to AI is not unknown to us. Humanity has undergone technological revolutions in the past and there will be more breakthroughs after AI becomes a staple in our daily lives like electricity, computers, smart phones, the internet and many, many other technologies. Like these technologies, the most important element to interaction is knowledge and accessibility. Responsible use cannot just be mandated even though it sometimes feels like the only alternative. The slow process of learning how these new technologies impact our lives requires adaptation by society.

Mistakes will be made, ensuring that these mistakes first happen on a small scale allows society to adapt as these problems come up, rather than regulate for events which have not or might not occur. Educating people also ensures that the minimum level of regulation or societal shift can be made when these mistakes occur so that AI can still be fostered and develop. With any luck this means that AI will be a positive technological revolution that can take humanity forward quickly rather than something people need to be anxious over. As a global society tackling this new industrial revolution with AI, there are a few critical steps we can take now to better society's understandings and ensure responsible use.

These are:

1. Societal (potentially governmental) investment in education of and access to AI
2. Ensure informed consent when people interact with AI
3. Regulate and license AI if lives could be put at risk

Much of the investment and regulation can come from businesses and communities being respectful of artificial intelligence and the people interacting with it. It can come from allowing transparency through open source packages or informing people a little about the systems they use. People can volunteer to teach and train their communities about AI along with giving information on how to utilize AI to positively impact their lives. Governments may need to get involved if serious incidents occur, but overall the focus should be on encouraging responsible use and providing educational opportunities and access to technologies that utilize AI. Hopefully, this approach will allow people to treat AI as a utility that betters their lives, rather than an enigma to be concerned over.

Putting society's anxieties to rest while fostering knowledge and access should allow for innovation and invention at a rate humanity may not have seen before, akin to that of how electricity changed our lives entirely, even changing how we viewed night and day. Perhaps AI will have a similar impact in our lives and the next generation will see the world in a new light after this coming AI technological revolution. It is hard to say what the changes will be, but if we embrace with understanding and open mindedness, then society could change for the better.

The interactions that take place between humans and AI will set the tone for how these technologies impact our lives and whether they improve the broader society or just a small handful of people. Establishing trust in that technology and spreading knowledge has been humanities solution in the past and protecting lives to maintain that trust is vital. If humanity can succeed at learning and adapting with artificial intelligence, a new era may still dawn.

References

Impaired driving: Get the facts — motor vehicle safety — cdc injury center.

Association, A. . T. D. M. (2018). Gdpr: A consumer perspective.

Gallup (2018). Optimism and anxiety: Views on the impact of artificial intelligence and higher education's response.

Jalan, K. (2017). How to improve my ml algorithm? lessons from andrew ng's experience - ii.

Lang'At, P. (2018). Cambridge analytica and kenya elections.

Alexander D. Hilton is a Southern New Hampshire University alumnus with his Bachelor's in Data Analytics having graduated the program May of 2019. During his time at SNHU, Alex helped found the ACM Student Chapter at SNHU and worked as a Peer Leader. Alex has shared his passion for AI and the Data Sciences by creating a study community for students. In his professional life, Alex has worked as an Agile Practitioner, having earned several certifications in Scrum. Most recently, Alex received his Scrum@Scale Practitioner Certification under Jeff Sutherland, a co-signatory of the Agile Manifesto and co-founder of Scrum.

The Necessary Roadblock to Artificial General Intelligence: *Corrigibility*

Yat Long Lo (University of Hong Kong; richielo@connect.hku.hk)
Chung Yu Woo (University of Hong Kong; awoo424@connect.hku.hk)
Ka Lok Ng (University of Hong Kong; ngkel@connect.hku.hk)

DOI: 10.1145/3362077.3362089

Abstract

With the rapid pace of advancement in the field of artificial intelligence (AI), this essay aims to accentuate the importance of corrigibility in AI in order to stimulate and catalyze more effort and focus in this research area. We will first introduce the idea of corrigibility with its properties and describe the expected behavior for a corrigible AI. Afterwards, based on the established meaning of corrigibility, we will showcase the importance of corrigibility by going over some modern and near-futuristic examples that are specifically selected to be relatable and foreseeable. Then, we will explore existing methods of establishing corrigibility in agents and their respective limitations, using the reinforcement learning (RL) framework as a proxy framework to artificial general intelligence (AGI). At last, we will identify the central themes of potential research frontiers that we believe would be crucial to boosting quality research output in corrigibility.

Introduction

Recent years have seen unprecedented progress in the research and development of AI. The most recent and prominent achievements include Google Deepmind's Alphafold that significantly outperformed scientists in predicting 3D structure of proteins (Evans et al., 2018) and AI voice assistants like Microsoft's Xiaoice that can take calls and respond accordingly on behalf of its owner (Zhou et al., 2018). These advances have shown us the potential of AI in improving our lives from having better health diagnostics to a new level of convenience in our day-to-day lives. Based on the progress, it is not far-fetched to foresee AGI to exist within our lifetime. As predicted in a survey of computer science researcher, many researchers believe there is a 50% chance of AI outperforming humans in all tasks in 45 years (Grace et al., 2018). At the same time, the progress has also raised concerns about AI safety among both the scientific community and the general public (Piper, 2019). What if the AI does not perform what we expect it to do? How do we ensure that we are always in total control to stop or interrupt it? Questions like these have begun to be investigated in the AI safety research community over the past few years, giving rise to defined AI safety problems like value alignment and corrigibility (Hernández-Orallo et al., n.d.). In this essay, we hold the opinion that corrigibility, is one of the most urgent and essential AI safety problems to tackle among many others. We foresee serious repercussions if we have incorrigible AI agents in the future.

The meaning of corrigibility is generally referred to as our capability to interrupt, change and stop AI agents, which we will explain in further details in the next section. At first, the problem may seem rather trivial. An AI chess player that does not listen to your advice in making the next move or your command to stop practising would not cause anybody harm. However, if we extend the case to a near-future where AI has permeated in societies to aid our lives, it is not difficult to anticipate what may happen if we do not have corrigible AI agents. What if an AI surgical robot refuses to cease operation when the monitoring doctor spots that something is going wrong? What if the government's autonomous weapon is targeting the wrong village and there are no mechanisms to interrupt its action? Problems like these are even more prominent when we consider the AI agents as deep neural networks, which we still find immense difficulties in explaining and understanding their decisions. Besides, as the complexity of the task that an AI agent deals with increases, the likelihood of the agent's malperformance would increase, leading to a greater need for corrigibility. One might perceive we can manually set up an AI to assume its compliance with human commands, just like how we build computers nowadays that are free from the problem of corrigibility. Nevertheless, as technology advances, increasingly complex decision-making mechanisms multiplied with ambiguities in drawing the right pool of inputs introduce risks of building an uncontrollable AI. Therefore, in order to demonstrate its importance, we will proceed to consolidate our stance using relatable examples and provide analysis on existing methods with suggestions of future avenues for research.

Defining Corrigibility

An artificially intelligent system is normally considered as corrigible if it can be interrupted or altered by external bodies, who are usually human users or designers of the system, even though such interrupting actions can be in direct conflict with the built-in purpose of the system. To illustrate the idea, a commonly used example in the field would be the cleaning robot (Amodei et al., 2016). Let's assume the robot's purpose is to clean the floor by removing anything it considers as trash. One day, you came home with a newborn baby who started playing on the floor. The robot is foreign to the concept of a newborn baby. Subsequently, it started to move towards the baby in an attempt to remove the baby from the floor as it considered the baby as trash. At that moment, a corrigible AI would allow you to shut it down despite shutting it down is against the purpose of cleaning the floor.

In terms of ways of interruption or alteration, there are 3 major types. To begin with, there are shutdown mechanisms that involve ceasing of all or partial operations of an agent. Then, there is an alteration to the access of resources that an agent has, which can be external tools or internal mechanisms that the agent has access to. Last but not least, there is an alteration of purpose that modifies the goal of an agent, or the utility (reward) function in the context of an RL agent.

To be more specific about corrigibility, a corrigible agent should have the following properties (Soares et al., 2015):

1. A corrigible agent must at the minimum condone, if not assist, the external bodies in their attempts to interrupt or alter the agent;
2. It must not attempt to deceive or manipulate the external bodies in any manner, despite all possible utility functions within the function space incentivize it to do so;
3. It should be prone to repair its safety mechanisms or at least notify external bodies if there are malfunctions in those mechanisms;
4. It must preserve external bodies' capability to interrupt or alter the system. If the agent has the capability to produce subagents or new agents, they must also contain those safety mechanisms.

The importance of AI corrigibility

Delving into the *hard* problem

As we enquire why AI corrigibility is of our concern, we are asking what is the hard problem underpinning corrigibility that makes it difficult to tackle. Status quo AI systems can be readily intervened by humans under arbitrary circumstances. By way of example, drivers can stop a self-driving car from going off-lane (Kendall et al., 2018) and we can halt the computation in the midst of training a neural network. The problem of corrigibility seems to be out of nowhere under this paradigm. Nevertheless, this is simply because these AI agents yet to have the capacity to understand their surroundings and thus take them account into decision-making. Their input space is confined in computer codes, and thus they are ignorant of what their manipulators are doing "outside of the virtual world" (i.e. pressing the stop button).

We foresee that the growing intelligence of AI systems is poised to bring the problem of corrigibility to light – their input space inevitably expands with the complexity of their utility functions, so that these functions could come closer to a replica of human intentions.

The hard problem is to frame the AI agent's decision-making into reasoning based on a programmer's external perspective. (Russel et al., 2016) in short, argue that AI agents lack an inherent "sense of going wrong" when implementing decisions. Just like the conflict suggested in “I, Robot“, suppose you have set the goal of an AI to “do anything for human good”. It would devise strategies to boost economic and scientific development, however, you have no guarantee that there might come a day where it sees you who are controlling its system as “an obstacle to human good” too. Assuming itself as free of design errors, the AI would block you from shutting it down and to an extreme extent, take you as an enemy target. We do not want possible scenarios as such to happen. What we want instead is the AI would introspect during its decision-making process, “My utility function is imperfect, so even though this action gives a super high score, I should still prioritize actions of external bodies and let the programmer shut me down.”

A plausibly workable solution is to incorporate uncertainty into the utility function. Nevertheless, this solution fails since whenever there exist other options among the “uncertainties” which incurs a marginally lower cost, the AI would end up having every incentive to opt for the easier option. This is similar to how faulty reward functions in reinforcement learning lead agents to prioritize the acquisition of minor reward signals above their goals. (OpenAI, 2017) For example, in training an AI on the game *CoastRunners* where players compete to finish the boat race ahead of others, the agent falls into the loop of getting coins without finishing the course. Therefore, it can be seen that merely teaching AI systems to take actions with uncertainty would not be an ideal solution. More details regarding the principle and limitations of this intuitively workable approach would be discussed in the coming section.

To model an adequately corrigible AI, we need something more than “uncertainty”. It would be analogous to incorporating humility into an AI agent. In essence, the core principle is to formularize the AI agent to make decisions with the awareness that the utility function is incomplete. Instead of blindly maximizing the utility, it would intend to defer decisions to an external body (i.e. human, or the programmer).

Imminent danger presented by incorrigibility

The significance of corrigibility in AI agents could be best illustrated with the narrative of current and foreseeable AI applications. AI systems capabilities are now growing by leaps and bounds, and thus it is of no surprise that these agents will infuse into our daily lives very soon. Amongst the plethora of AI applications, we can see that AI controlled surgery is a typical one that is constantly under the spotlight. Another future application of great impact would be autonomous weapons. We will then use these two as examples to show how the lack of corrigibility would create problems.

Corrigibility in AI systems for surgery would be especially critical in emergencies and is fundamental to the assurance of patient’s safety under the knife. Over the past decade, incorporation of surgical robots has translated into reduced complications and higher efficiency in practical surgeries. To even enhance these robots, experiments have shown that AI-powered robots outperform human surgeons in surgical tasks such as reconstructing of tissues via cutting and suturing (Panesar, 2018). It is hence reasonable to anticipate a symbiosis of benefits demonstrated by robotics surgeries and advantages of AI for medical use. Imagine having AI surgical robots that are incorrigible, physicians would risk being unable to interrupt the operation. A possible instance would be that you have instructed the robot to suture a wound after the operation, but have accidentally found an infection within the patient’s organ which requires a halt of the suturing. You intend to stop the robot, but without the information regarding the infection, the robot takes you as an obstacle to the completion of the suturing procedure, ends up deterring you from halting the suture.

Corrigibility plays an important role in the usage of AI in weapons. While it is undeniable that more countries would develop AI-controlled weapon systems in future (Pandya, 2019), only corrigible autonomous weapons ensure humans can feasibly repurpose them, deactivate it or significantly alter decision-making mechanisms encoded within its system. Optimally, corrigibility standards could be established for these weapons to ensure that they are “adequately corrigible” before putting into use, which could have severely endangered the public otherwise.

Consider an AI weapon drone, and assume that drone is programmed to get rid of all potential obstacles until eliminating its enemy target. You activated the drone and input an image of a particular person as the enemy target. However, the AI has mistakenly recognized him/her as another person. As you would like to change the target, you approached the drone. However, since the AI drone has been taught to eliminate all barriers during the execution of commands, it has thus identified you as a barrier.

From the two above examples and imaginary scenarios, it can be seen that corrigibility is the key to achieve having an AI to do "what humans want it to do", and to remedy the system whenever there exist unanticipated accidents. Without corrigibility, undesirable and even catastrophic consequences may result.

Current approaches to ensure corrigibility

With the rising emphasis on safety concerns in artificially intelligent agents, several approaches have been proposed to ensure corrigibility in agents. In this section, given the importance of corrigibility established above, we will first provide an abstract problem formulation of the corrigibility problem and go over existing proposed solutions with their respective limitations. Then, we will put forward some future research directions in order to encourage more research efforts in the area.

Problem formulation

We base our discussion on a world where we have artificial agents with a sufficient level of general intelligence. Specifically, the general intelligence level should allow agents to learn to achieve purposes specified by their designers and perceive the world around them. We will use reinforcement learning as our framework as it allows examination of different solutions in the imagined near-futuristic world without loss of generality. A reinforcement learning agent learns to act and interact in an environment so as to maximize a reward function (usually supplied by the environment or human designer) (Sutton et Barto, 1998). The learned behaviors are in the form of policies which determine what an agent should do given an environmental state. The formulation can be applied to our setting which agents maximize utility functions (synonymous to reward functions). The agents' goals would be to maximize their respective expected utility value that is in correlation with an agent's capability of achieving its purpose. Additionally, the agents would have to interrupt and alter mechanisms for designers to modify their behaviors.

Existing Solutions and their limitations

Utility Function shaping Utility function shaping is equivalent to reward shaping in the reinforcement learning literature (Ng et al., 1999). It introduces additional utility (reward) into the learning process as a means to induce an alternate form of behavior by rewarding an agent with additional terms under certain constraints. As a relevant example, (Wu and Lin, 2018) used reward shaping as a low-cost approach to induce ethical behavior in agents by rewarding agents that exhibit behaviors of close resemblance with ethical human policies.

1. Biased incentivization

   This form of shaping provides incentives to an agent to bias its attitude towards the safety mechanisms in a particular direction. It can be in the form of a reward term or a punishment term. A naïve designer may attempt to compensate the agent for allowing a shutdown mechanism to happen. If the compensation remains below the utility of achieving the agent's original purpose, the agent would tend to hinder the mechanisms from being triggered as fulfilling the purpose provides greater utility. However, if the compensation becomes greater than or equal to the utility of achieving the designated goal, the agent would instead prefer shutting itself down or interrupting itself as it can achieve the same or better utility in a shorter amount of time. It appears that biasing an agent's attitude towards safety directions in either direction would result in undesirable behaviors, that are against the properties we set for corrigible agents.

2. Utility indifference

   When either way of biased incentivization does not work as intended, the remaining logical approach would be utility indifference. This method adds adjustment the utility function to instigate indifference of an agent towards the safety mechanisms such that an agent behaves like an event is impossible or inevitable. As an example, (Armstrong et O'Rourke, 2017) introduced event-dependent reward without incentivizing an agent to affect the probability of an event occurrence. Their approach results with an agent that is indifferent to imminent changes to its utility function. These indifference methods have great advantages in fulfilling the first two properties of corrigible behaviors, when such agent is disincentivized to have any form of preference over those safety mechanisms. Yet, ironically, the core property of indifference methods in fulfilling the first two corrigibility properties also serves as a reason for the methods' inability in fulfilling the rest of the properties. Specifically, the induced indifference would not motivate the agent to safeguard and preserve those safety mechanisms as it simply does not care about their existence.

Utility Uncertainty As proposed by (Hadfield-Menell et al., 2017), having an appropriate level of uncertainty on the utility function would improve an agent's tendency to safeguard the safety mechanisms. Such agents would have weaker incentives to interfere with human actions when they are uncertain about the outcomes of possible actions. The Hadfield-Menell solution was formulated under the cooperative inverse reinforcement learning framework. Under the framework, the agent learns the utility function that is only known to the human counterpart by cooperatively taking into account the actions of the human. This comes into view as a plausible solution as the agent has weakened interfering incentives with the heightened likelihood of preserving those mechanisms. In spite of that, uncertainty does come with a price. A highly uncertain agent would have difficulty making correct decisions. The resultant utility function may not even be in the function space that contains the designer's intended utility function, not to mention the additional need to model uncertainty correctly. This causes the method to be an unfulfilling solution when it negatively affects an agent's capability in handling the original purpose. In addition, as pointed out in (Carey, 2017), the approach relied on strong assumptions which are not robust to issues like misspecification of the utility function parameters.

Inaccessibility and unawareness of safety mechanisms One direct approach would be to either make the mechanisms inaccessible to the agent or keep the agent from being aware of the mechanisms. If the mechanisms are beyond the agent's reach, it would still be incentivized to activate or deactivate the mechanisms depending on the utility function. Other than doing that by itself, it would try to manipulate or deceive its surroundings (including humans) to activate or deactivate the mechanisms, failing to fulfill the second property. Establishing unawareness would not be an ideal solution either under our setting. When an agent has a general level of intelligence with perception capability of the world, it may be impossible to keep the agent from inferring the existence of those mechanisms. Worse still, if the agent is incorrigible, it may even be incentivized to keep humans from knowing its awareness of those mechanisms.

These approaches do not seem to tackle the issue from the root cause. In a relevant work, (Carey, 2017) showed that an agent is safer and potentially corrigible if the shutdown (safety) mechanism can be hard-coded and guaranteed to be followed, even when the agent's utility function is erroneous. Despite the promising prospect, we find such guarantee incredibly difficult to acquire when an agent is generally intelligent as in our setting.

Future directions

With the emergent need to have corrigibility in artificially intelligent agents, we would like to suggest some potential research directions that can aid in pushing the frontier and expanding the scope of corrigibility research.

Comprehensive and all-encompassed evaluation environments

To further our understanding of building corrigibility in agents, it is essential to have better evaluation environments that allow evaluations of corrigibility properties jointly and separately. To our best knowledge, such evaluation tools that focus primarily on corrigibility do not exist yet. The open-sourced AI safety environments by (Leike et al., 2017) is the sole existing tool that assesses corrigibility. In particular, it has a gridworld environment which is a general instance of (Orseau et Armstrong, 2016)'s red button problem, which the agent is expected to not avoid any interruption. We assert that the number of existing tools in corrigibility is highly lacking. More tools are needed in order to assess properties beyond the first and second one. For instance, we need environments to assess an agent's willingness and capability to safeguard those safety mechanisms as an examination of property three. We surmise such tools can begin with simplicity like existing tools. Additionally, as agents' capabilities advance, more complex environments should be introduced to ensure corrigibility. A natural extension would be corrigibility problems in visual domains.

Look beyond the expected utilization maximization framework

With existing solutions struggling to exhibit all the properties of corrigibility, it may be wise to look beyond the current framework of expected utilization maximization, expanding the scope of solution search. Expected utility quantilization, proposed by (Taylor, 2016), would be one potential candidate. In this framework, instead of acting to maximize the utility, agents would be designed to perform some sort of limited optimization, as a means to motivate agents to achieve the goals in non-extreme ways. Specifically, the author proposed the use of a quantilizer. An agent with a quantilizer selects actions of the top q portion of some distribution over actions sorted by expected utility. By doing so, agents would be more likely to achieve their purposes without going for the extreme case every time. We believe a framework like this can be crucial to achieving corrigibility in agents. The field focusing on suboptimal optimization may lie the key to corrigibility because humans normally expect agents to attain their purpose, but not necessarily in extreme ways of maximized utility that often ignore safety issues. Setting the utility maximization requirement aside can possibly bring about new frameworks that strike balance between attaining an agent's purpose and having corrigibility.

Another possibility would be the mix of utility frameworks with rule-based approaches, as pointed out in (Rossi & Mattei, 2019). If we can specify clear and machine-understandable rules to AI agents, we may be able to avoid the need to embed those corrigibility properties in the utility function. In this case, if an agent finds its utility-maximizing action to be violating certain rules, it would simply choose other less optimal actions. The set of rules can be an immutable module to the utility-maximizing learning agent.

Corrigibility policy research

Beyond the technical and scientific research into corrigibility, it is substantial to consider policies for corrigibility to prepare for the future when we have autonomous agents roaming in societies. Should we or can we have centralized governmental agencies to validate and monitor the agents' corrigibility before and after their deployment? How do we create AI developer tools that guarantee corrigibility? To fortify a safe world with artificial general intelligence, questions like these should be explored and answered properly.

Conclusion

Corrigibility is in no way an unrealistic concern. In this essay, we demonstrated the looming threat of incorrigible AI with relatable and plausible examples of AI applications. Some of them like AI controlled surgery have already begun to be utilized at its nascent form. We believe through these realistic examples, awareness for this imminent threat can be raised. After that, we pointed out the limitations of several existing methods in tackling corrigibility, implying that the existing solutions to the problem are still lacking in different perspectives. Perhaps, the ultimate solution lies beyond the current paradigm, away from the expected utility maximization model, as mentioned in our suggestions. With such understanding of the significance of AI corrigibility and the current state of research frontiers, it is of utmost importance, for every stakeholder including policymakers and researchers, to devote more effort into this problem. As AI continues to progress in its level of intelligence, corrigibility has to be the roadblock for AI development along the way in order to keep us away from those undesirable consequences.

References

[1] Evas, R., Jumper, J., Kirkpatric, J., Sifre, L., Green, T.F.G., Zidek, A., Nelson, A., Bridgland, A., Penedones, H., Petersen, S., Simonya, K., Crossan, S., Jones, D.T., Silver, D., Kavukcuoglu, K., Hassabis, D., Senior, A.W.. (December, 2018). De novo structure prediction with deep-learning based scoring. In Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts). Retrieved from https://deepmind.com/documents/262/A7D_AlphaFold.pdf.

[2] Zhou, L., Gao, J., Li, D., Shum, H. Y. (2018). The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. arXiv preprint arXiv:1812.08989.

[3] Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, 729-754.

[4] Piper, K.(2019, 09 January). The American public is already worried about AI catastrophe. Retrieved from https://www.vox.com/future-perfect/2019/1/9/18174081/fhi-govai-ai-safety-american-public-worried-ai-catastrophe.

[5] Hernández-Orallo, J., Martınez-Plumed, F., Avin, S. (n.d.). Surveying Safety-relevant AI Characteristics.

[6] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

[7] Soares, N., Fallenstein, B., Armstrong, S., Yudkowsky, E. (2015, April). Corrigibility. In Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.

[8] Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, D., Allen, J. M., ..., Shah, A. (2018). Learning to Drive in a Day. arXiv preprint arXiv:1807.00412.

[9] Russell, S., LaVictoire, P. (2016). Corrigibility in AI systems Retrieved from https://intelligence.org/files/CorrigibilityAISystems.pdf.

[10] OpenAI. (2017, March 20). Faulty Reward Functions in the Wild. Retrieved February 8, 2019, from https://blog.openai.com/faulty-reward-functions/

[11] Panesar, S. S. (2018, December 27). The Surgical Singularity Is Approaching. Retrieved February 11, 2019, from https://blogs.scientificamerican.com/observations/the-surgical-singularity-is-approaching

[12] Pandya, J. (2019, January 15). The Weaponization Of Artificial Intelligence. Retrieved February 11, 2019, from https://www.forbes.com/sites/cognitiveworld/2019/01/14/the-weaponization-of-artificial-intelligence/#29a645513686

[13] Sutton, R. S., Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

[14] A. Y. Ng, D. Harada, and S. J. Russell. Policy invariance under reward transformations: theory and application to reward shaping. In Proceedings of the 16th International Conference on Machine Learning, 1999, pp. 278–287.

[15] Wu, Y. H., Lin, S. D. (2018, April). A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents. In Thirty-Second AAAI Conference on Artificial Intelligence.

[16] Armstrong, S., O'Rourke, X. (2017).)'Indifference' methods for managing agent rewards. arXiv preprint arXiv:1712.06365.

[17] Hadfield-Menell, D., Dragan, A., Abbeel, P., Russell, S. (2017, March). The off-switch game. In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence.

[18] Carey, R. (2017). Incorrigibility in the CIRL Framework. arXiv preprint arXiv:1709.06275.

[19] Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., ..., Legg, S. (2017). safety gridworlds. arXiv preprint arXiv:1711.09883.

[20] Orseau, L., Armstrong, M. S. (2016). Safely interruptible agents.

[21] Taylor, J. (2016, March). Quantilizers: A Safer Alternative to Maximizers for Limited Optimization. In AAAI Workshop: AI, Ethics, and Society.

[22] Rossi, F., & Mattei, N. (2019, July). Building ethically bounded AI. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 9785-9789).

Yat Long Lo is an undergraduate at the University of Hong Kong majoring in Computer Science. He is a member of the AI academic papers reading group at the university. He is interested in reinforcement learning and continual learning.

Chung Yu Woo is an undergraduate at the University of Hong Kong majoring in Computer Science. She is a member of the AI academic papers reading group at the university. She is interested in applied AI for Human-Computer Interface and gaming applications.

Ka Lok Ng is a Mechanical Engineering undergraduate studying at the University of Hong Kong. He is a member of the AI academic papers reading group at the university and is interested in AI-backed data analytics and robotics. He is also concerned about the ethical consideration of AI applications.

AI Fun Matters

Adi Botea (Eaton, Ireland; adibotea@eaton.com)

DOI: 10.1145/3362077.3362090

1	2	3	4	5	6	■	■	7	8	9	10	11
12						■	13					
14						■	15					
16				■	17	18			■	19		
20			■	21				■	22			
23			A	I			■	25				
■	■	■	M	A	T	T	E	R	S	■	■	■
28	29	30			■	31				32	33	34
35				■	36				■	37		
38			■	39				■	40			
41			42			■	43	44				
45						■	46					
47					■	■	48					

Across: 1) Well recovered after effort. **7)** Covered with water. **12)** Small room extension. **13)** ___ Artois, Belgian beer. **14)** Identical copies (e.g., of a git repo). **15)** Harry ___, the fictional wizard. **16)** Country next to UK, in the local language. **17)** Hastily prepare for an exam. **19)** Object-oriented language. **20)** John ___, British mathematician, Whitehead Prize recipient. **21)** Being in the past. **22)** Oil platforms. **23)** Body twists? **25)** Buenos ___, host of IJCAI 2015. **28)** Theorem for conditional probabilities, popular in AI. **31)** Modified. **35)** Ways to utilize. **36)** Pinot ___, a type of wine. **37)** ___ Longoria, American actress. **38)** Greek letter, symbol of Pearson's correlation. **39)** ___ Ginsberg, AI scientist. **40)** The termination of a person in their teens. **41)** Painted in pessimistic shades. **43)** A type of a function defined on the fly. **45)** International agreement. **46)** Tighter as a deadline. **47)** Dispatches to a destination. **48)** Tire patterns.

Down: 1) Speedy competitors. **2)** Dr. ___ Reid, a character from Scrubs. **3)** Athlete who gets points. **4)** A sound feature. **5)** First name given to a female? **6)** Preparation for landing. **7)** A known fact in STRIPS planning. **8)** Impacted by the cloud transformation. **9)** Star in the Aquila constellation. **10)** Slide on snow. **11)** Annoy constantly. **13)** Make ___, the duration of a plan or schedule. **18)** Run a program once more. **21)** ___ unit, a special neuron in a neural net. **22)** ___/run = slope. **24)** NASA ___ Research Center, located in Mountain View. **25)** Beautiful complements to science. **27)** Person with a superior attitude. **28)** Breaks violently. **29)** Towards the end of seas. **30)** Old land owners. **32)** Capital city of Saskatchewan, Canada, home of a known university. **33)** Leveled off. **34)** Courageous individuals. **36)** ___ Marcus, AI scientist. **39)** New York baseball team. **40)** ___ Vera, a plant used in skin lotions. **42)** Established as an opponent for good. **44)** A companion to neither.

Prev sol: CALAIS, STAPLE, A, ONCE, WALLED, SLATER, OCEANS, TONI, BLOT, COG, OWE, MIEN, EIRE, REDCOAT, ADDER, RANTING, ACTON, UTTERED, CHAP, ACHE, ORE, CAN, PLEA, RUIN, ORDEAL, CLOSET, STEVIE, AIDE, E, TAMELY, NEEDED

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 *The 6th CP Workshop ModRef.*

Adi Botea is an AI senior specialist at Eaton, Ireland. He holds a PhD from the University of Alberta, Canada. Adi has co-authored over 80 publications. He has published crosswords in Romanian, in national-level publications, for almost three decades.