AI Matters

Annotated Table of Contents

Full article: http://doi.acm.org/10.1145/3137574.3137585

ACM SIGAI Student Essay Contest Winner

The Ethics of Automated Behavioral Microtargeting

Dennis G. Wilson

Full article: http://doi.acm.org/10.1145/3137574.3139451

ACM SIGAI Student Essay Contest Winner

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: 10.1145/3137574

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.

Submit to AI Matters!

We're accepting articles and announcements now for future issues. 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, Volume 3, Issue 3

Eric Eaton, Co-Editor (University of Pennsylvania; aimatters@sigai.acm.org)
Amy McGovern, Co-Editor (University of Oklahoma; aimatters@sigai.acm.org)
DOI: 10.1145/3137574.3137575

Welcome to the third issue in our third year of *AI Matters*! In the spring, ACM SIGAI sponsored a student essay contest on the "Responsible Use of AI Technologies." This issue features the first set of winning essays from the contest, with the second set of winning essays to appear in the next issue.

In addition to having their essay appear in *AI Matters*, the contest winners received either monetary prizes or one-on-one Skype sessions with leading AI researchers. The students reported exceptional experiences:

> I just had an absolutely phenomenal conversation with Eric Horvitz. Really so spectacular. Thank you so so so much for enabling this to happen. .... This was such a privilege and wonderful experience.

> It was great talking to [Peter Norvig] and we discussed a broad range of topics from Autonomous Weapons, changes in the job market caused by automation to AI techniques to help people programming. It was very inspiring to talk to him!

Speaking of leading researchers, this issue's **AI Interviews** column highlights Maja Matarić, the Vice Dean for Research and the Director of the Robotics and Autonomous Systems Center at the Univ. of Southern California.

In AI news, this issue includes the 2016-2017 ACM SIGAI Activity Report, which summarizes the annual activities of the SIGAI, reflections on *The 50 Years of the ACM Turing Award Celebration* by two student SIGAI scholars, and a report on the new China chapter of SIGAI.

The (buzz)word of this issue is *multi-agent path finding*. "What's that?" you say. Well, you can read all about it in the **AI Buzzwords Explained** column. Or, if deep learning is more your thing, then check out the **AI Education** column on resources for teaching and learning about deep neural networks.

Thanks for reading! Don't forget to send your ideas and future submissions to *AI Matters*!

Eric Eaton is a Co-Editor of AI Matters. He is a faculty member at the University of Pennsylvania in the Department of Computer and Information Science, and in the General Robotics, Automation, Sensing, and Perception (GRASP) lab. His research is in machine learning and AI, with applications to robotics, sustainability, and medicine.

Amy McGovern is a Co-Editor of AI Matters. She is an Associate Professor of computer science at the University of Oklahoma and an adjunct associate 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.

Submit to AI Matters!

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

AI Events

Michael Rovatsos (University of Edinburgh; mrovatso@inf.ed.ac.uk)
DOI: 10.1145/3137574.3137576

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

International Conference on Computational Approaches to Diversity in Interaction and Meaning

San Servolo/Venice, Italy, October 7-9, 2017
www.essence-network.com/essence-final-conference
Diversity-awareness, understood as the capability of an intelligent system to take the heterogeneity of human or artificial agents it is interacting with into account when making decisions, has been researched by many communities across several disciplines in the past, including semantic technologies, multiagent systems, knowledge representation and reasoning, and NLP. This conference will provide a forum for leading experts from the above (and other) areas to discuss key research issues surrounding diversity-aware AI. Participation is by invitation only, and financial support is available – please contact essence-info@inf.ed.ac.uk if you are interested in attending.

32nd International Conference on Automated Software Engineering

Urbana-Champaign, USA, October 30 to November 3, 2017
ase2017.org
The IEEE/ACM Automated Software Engineering (ASE) Conference series 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. ASE 2017 features high quality contributions describing significant, original, and unpublished results.

9th International Joint Conference on Computational Intelligence (IJCCI 2017)

Funchal, Portugal, November 1-3, 2017
www.ijcci.org
The purpose of the International Joint Conference on Computational Intelligence (IJCCI) is to bring together researchers, engineers and practitioners interested on the field of Computational Intelligence both from theoretical and application perspectives. Four simultaneous tracks will be held covering different aspects of Computational Intelligence, including evolutionary computation, fuzzy computation, neural computation and cognitive and hybrid systems. The connection of these areas in all their wide range of approaches and applications forms the International Joint Conference on Computational Intelligence.

CRA Summit on Technology and Jobs

Washington DC, USA, December 12, 2017
cra.org/events/summit-technology-jobs/
The goal of the summit is to put the issue of technology and jobs on the national agenda in an informed and deliberate manner. The summit will bring together leading technologists, economists, and policy experts who will offer their views on where technology is headed and what its impact may be, and on policy issues raised by these projections and possible policy responses. The summit is hosted by the Computing Research Association, as part of its mission to engage the computing research community to provide trusted, non-partisan input to policy thinkers and makers.

10th International Conference on Agents and Artificial Intelligence (ICAART 2018)

Funchal, Portugal, January 16-18, 2018
www.icaart.org
The purpose of the International Conference on Agents and Artificial Intelligence 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.

23rd International Conference on Intelligent User Interfaces (IUI 2018)

Tokyo, Japan, March 7-11, 2018
iui.acm.org
ACM IUI 2018 is the 23rd annual meeting of the intelligent interfaces community that will be held in Tokyo, on March 7th-11th, 2018. IUI is where the HCI and AI communities meet, and the focus of the conference is to improve the interaction between humans and machines, by leveraging both traditional HCI approaches and solutions that involve state-of-the art AI techniques like machine learning, natural language processing, data mining, knowledge representation and reasoning. Along with 25 topics in AI and HCI, this year IUI especially encourages submissions on explainable intelligent user interfaces.
Submission deadline: October 8, 2017

20th International Conference on Enterprise Information Systems (ICEIS 2018)

Funchal, Portugal, March 21-24, 2018
www.iceis.org
The purpose of the 20th International Conference on Enterprise Information Systems (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: October 18, 2017.

31st International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE-2018)

Montreal, Canada, June 25-28, 2018
ieaaie2018.encs.concordia.ca
IEA/AIE 2018 continues the tradition of emphasizing 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. IEA/AIE-2018 will include oral presentations, invited speakers, and special sessions.
Submission deadline: November 27, 2017

17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)

Stockholm, Sweden, July 10-15, 2018
celweb.vuse.vanderbilt.edu/aamas18
AAMAS is the largest and most influential conference in the area of agents and multiagent systems. The aim of the conference is to bring together researchers and practitioners in all areas of agent technology and to provide a single, high-profile, internationally renowned forum for research in the theory and practice of autonomous agents and multiagent systems. AAMAS is the flagship conference of the non-profit International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS). This edition will be a part of the Federated AI Meeting (FAIM) where AAMAS is co-located with several AI conferences including the International Joint Conference on Artificial Intelligence (IJCAI)/European Conference on Artificial Intelligence (ECAI) and the International Conference on Machine Learning (ICML).
Submission deadline: November 14, 2017.

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

ACM SIGAI Activity Report

Sven Koenig (elected; ACM SIGAI Chair)
Sanmay Das (elected; ACM SIGAI Vice-Chair)
Rosemary Paradis (elected; ACM SIGAI Secretary/Treasurer)
Eric Eaton (appointed; ACM SIGAI Newsletter Editor-in-Chief)
Yolanda Gil (appointed; ACM SIGAI Past Chair)
Katherine Guo (appointed; ACM SIGAI Membership and Outreach Officer)
Bojun Huang (appointed; ACM SIGAI Information Officer)
Albert Jiang (appointed; ACM SIGAI Education Officer)
Benjamin Kuipers (appointed; ACM SIGAI Ethics Officer)
Nicholas Mattei (appointed; ACM SIGAI Ethics 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)
Plamen Petrov (appointed; ACM SIGAI Industry Liaison Officer)
Michael Rovatsos (appointed; ACM SIGAI Conference Coordination Officer)
David Stork (appointed; ACM SIGAI Award Officer)
DOI: 10.1145/3137574.3137577

Abstract

We are happy to present the annual activity report of ACM SIGAI, covering the period from July 2016 to June 2017.

The scope of ACM SIGAI consists of the study of intelligence and its realization in computer systems (see also its website 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 of ACM SIGAI come from academia, industry and government agencies worldwide. ACM SIGAI is proud of the fact that many AI researchers in the past year received ACM honors, including becoming ACM fellows as well as receiving other awards.

ACM SIGAI is committed to increase its activities in order to support its members even better. In order to do so, ACM SIGAI has created several new appointed officer positions, namely an industry liaison officer, an award officer and two ethics officers.

In the course of the last year, ACM SIGAI has been responsive to specific events and circumstances as well as continued to support and expand a range of regular activities.

Responsive Initiatives in the Last Year

ACM SIGAI actively supported the founding of a new ACM SIGAI chapter in China with help from the membership and outreach officer. ACM SIGAI China held its first event, the ACM SIGAI China Symposium on New Challenges and Opportunities in the Post-Turing AI Era, as part of the ACM China Turing 50th Celebration Conference on May 12-14, 2017 in Shanghai.

ACM SIGAI held the ACM SIGAI Student Essay Contest on the Responsible Use of AI Technologies (run by one of the ethics officers), where students could win five cash prizes of US$500 or skype conversations with five very senior AI researchers from academia or industry if their essay provided good answers to the following two questions:

- What do you see as the 1-2 most pressing ethical, social or regulatory issues with respect to AI technologies?
- What position or steps can governments, industries or organizations (including ACM SIGAI) take to address these issues or shape the discussions on them?

The winning essays will be published in the ACM SIGAI newsletter.

ACM SIGAI extended its coordination and collaboration with a variety of groups, both inside and outside of ACM:

- ACM SIGAI started to participate in the ACM US Public Policy Council (USACM). USACM 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. The public policy officer, for example, facilitated talks between the leaderships of USACM and the American Association for the Advancement of Science (AAAS) on areas of potential collaboration.
- ACM SIGAI started to participate in the IEEE Global Initiative for Ethical Considerations in AI and Autonomous Systems. The purpose of this initiative is to ensure that every technologist is educated, trained and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems.
- ACM SIGAI also provided a response to the public request for information from the US Office of Science and Technology Policy (OSTP) in 2016 on Preparing for the Future of AI, thus supporting the US government in making decisions concerning AI technologies and their applications.

In response to developments regarding international travel policies, ACM SIGAI released the following public policy statement in 2017 via an effort of the public policy officer:

> "The ACM SIGAI executive committee shares the view of its parent organization that 'the open exchange of ideas and the freedom of thought and expression are central to the aims and goals of ACM. ACM supports the statute of International Council for Science in that the free and responsible practice of science is fundamental to scientific advancement and human and environmental wellbeing. Such practice, in all its aspects, requires freedom of movement, association, expression and communication for scientists. All individuals are entitled to participate in any ACM activity.' SIGAI is working on policies to support inclusive participation in our AI-related activities. We encourage event organizers to share their efforts and experiences with us through our AI Matters newsletter at `aimatters@sigai.acm.org` and blog postings at `sigai.acm.org/aimatters/blog/`."

ACM SIGAI also actively supported the Journal of Human-Robot Interaction (JHRI) in its desire to become an ACM journal and be included in the ACM Digital Library. JHRI will become the ACM Transactions on Human-Robot Interaction in January 2018.

Continuing Activities

Organizing events:

ACM SIGAI organized the annual ACM SIGAI Career Network Conference (CNC), overseen by the vice chair. CNC showcases the work of early career researchers (including students) to their potential mentors and employers. Each early career researcher is mentored by a senior AI researcher, with small group mentoring sessions as well as individual advice. CNC 2016 was held at Northeastern University in Boston, Massachusetts, on October 19-20, 2016 (and ACM SIGAI gratefully acknowledges the support and hospitality of Northeastern University). 36 early career researchers presented their work via talks and posters, and two panels (on Career Options and Getting Started, featuring senior AI researchers and practitioners from academia, industry and the public sector) completed the program.

ACM SIGAI has an agreement with the Association for the Advancement of AI (AAAI) to jointly organize 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. This event was held at AAAI 2017 as planned.

Supporting international conferences and other events:

ACM SIGAI processed requests for co-sponsored and in-cooperation status from 27

conferences. The conference coordination officer improved the support provided to conference organizers by contacting them personally immediately after approval, inviting them to publicize their conference via the ACM SIGAI newsletter and mailing lists, and following up with a request for a conference report after the conference, in order to publish it in the ACM SIGAI newsletter and blog.

ACM SIGAI has an agreement with AAAI to co-sponsor, jointly with AAAI, the annual joint doctoral consortium at the AAAI conference. The doctoral consortium 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 that act as their mentors.

ACM SIGAI also co-sponsored the following conferences (*future events are in italics*):

- ACM/IEEE International Conference on Human-Robot Interaction (HRI 2017)
- 22nd International Conference on Intelligent User Interfaces (IUI 2017)
- *ACM/IEEE International Conference on Human-Robot Interaction (HRI 2018)*
- *ACM/IEEE International Conference on Automated Software Engineering (ASE 2017)*
- *ACM/IEEE International Conference on Automated Software Engineering (ASE 2018)*
- *International Conference on Web Intelligence (WI 2017)*
- *International Conference on Web Intelligence (WI 2018)*
- *23rd International Conference on Intelligent User Interfaces (IUI 2018)*
- *24th International Conference on Intelligent User Interfaces (IUI 2019)*

In addition, ACM SIGAI granted in-cooperation status to the following conferences:

- Swarm/Human Blended Intelligence Workshop (SHBI 2016)
- 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)
- International Conference on Agents and Artificial Intelligence (ICAART 2017)
- International Knowledge System Conference (KMIKS 2017)
- 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)
- 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)
- International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE 2017)
- *14th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2017)*
- *11th ACM Conference on Recommender Systems (RecSys 2017)*
- *International Conference on the Foundations of Digital Games (FDG 2017)*
- *International Joint Conference on Rules and Reasoning (RuleML+RR 2017)*
- *4th International Workshop on Sensor-based Activity Recognition and Interaction (iWOAR 2017)*
- *Data Institute San Francisco Conference (DSCO 2017)*
- *9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2017)*
- *9th International Joint Conference on Computational Intelligence (IJCCI 2017)*
- *10th International Conference on Agents and Artificial Intelligence (ICAART 2017)*
- *7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*
- *11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018)*
- *31st International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE 2018)*
- *12th ACM Conference on Recommender Systems (RecSys 2018)*
- *31th Annual ACM Symposium on User Interface Software and Technology (UIST 2018)*
- *20th International Conference on Enterprise Information Systems (ICEIS 2018)*

ACM SIGAI has an agreement with the International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS) to sponsor the ACM SIGAI Autonomous Agents Research Award. The ACM SIGAI Autonomous Agents Research Award is 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 2017 ACM SIGAI Autonomous Agents Research Award was presented at AAMAS 2017 to David Parkes from Harvard University for his work on a variety of topics in multi-agent systems and economics.

Increasing the visibility of AI research:

ACM SIGAI 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 general interest to the computer science research community to the Research Highlight Track. This way, ACM SIGAI helps to make important AI research results visible to many computer scientists.

Supporting student members:

ACM SIGAI believes that funding students is a good way to ensure vitality in the AI community and thus a good investment in the future. Consequently, it awarded a number of scholarships to students to attend conferences co-sponsored by it as well as the 50 Years of the ACM Turing Award Celebration. 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. ACM SIGAI changed the period of time before students who join ACM SIGAI can apply for financial benefits. There is now a 3-month membership requirement before students can apply for fellowships and travel support.

Communicating with and supporting members:

ACM SIGAI publishes four issues of its newsletter AI Matters per year. AI Matters features articles of general interest to the AI community, from research overview articles to dissertation abstracts. The editors-in-chief instituted a number of reforms over the last year. For example, they started a number of recurring columns in an effort to focus on the needs and interests of individual populations of the membership and promote content creation for each issue. These columns are led by individual column editors, who are responsible for soliciting content or writing the column each quarter. These columns have included

- AI Interviews (with interesting people from academia, industry and government),
- AI Amusements (including AI humor, puzzles and games),
- AI Education (led by one of the education activities officers),
- AI Policy Issues (led by the public policy officer),
- AI Buzzwords (which explains new AI concepts or terms),
- AI Events (which includes conference announcements and reports),
- AI Dissertation Abstracts and
- News from AI Groups and Organizations.

To promote readership, the editors-in-chief have moved to an open-access model (where AI Matters is openly available on the ACM SIGAI website) and instituted a new AI Matters blog (at `sigai.acm.org/aimatters/blog/`) to feature timely posts and promote discussion among community members. For example, the 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. Behind the scenes, the editors-in-chief have also revamped the editorial process and created tools to streamline the assembly of each issue.

ACM SIGAI organized or co-organized ACM webinars to inform ACM members about AI topics, such as the Town Hall on AI, Machine Learning, and More in 2016 (run by the secretary/treasurer) and the Panel and Town Hall on Big Thoughts and Big Questions about Ethics in AI (run by one of the ethics officers). The webinars were streamed live but the videos are archived at `learning.acm.org/webinar/`.

Additional ACM SIGAI membership benefits include reduced registration fees at many of its 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 is working on increasing the communication with its current membership (for example, via social media and membership

surveys) and intensifying its activities that reach students (for example, via the student essay contest) and industry professionals (for example, via the webinars). Additional activities in this directions are currently being planned. ACM SIGAI also intends to continue expanding the number of co-sponsored and in-cooperation conferences and its efforts to influence public policy and further the discussion on the responsible use of AI technologies. Finally, it 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.

To help further the ACM SIGAI's activities, you should consider becoming a SIGAI member!

For details, see http://sigai.acm.org/.
The SIGAI mailing list is open to all.

AI Profiles: An Interview with Maja Matarić

Amy McGovern (University of Oklahoma; amcgovern@ou.edu)
Eric Eaton (University of Pennsylvania; eeaton@cis.upenn.edu)
DOI: 10.1145/3137574.3137578

Abstract

This column is the fourth in our series profiling senior AI researchers. This month we interview Maja Matarić.

Introduction

Our fourth profile for the interview series is Maja Matarić, Vice Dean for Research and the Director of the Robotics and Autonomous Systems Center at the University of Southern California.

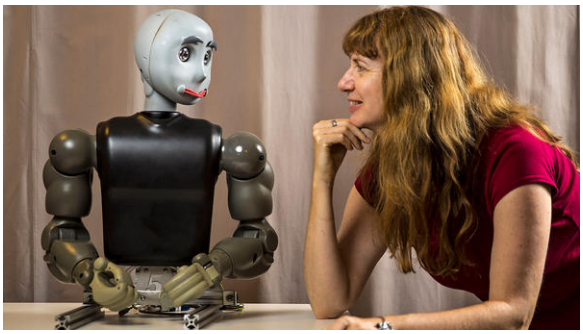


Figure 1: Maja Matarić

Biography

Maja Matarić is professor and Chan Soon-Shiong chair in Computer Science Department, Neuroscience Program, and the Department of Pediatrics at the University of Southern California, founding director of the USC Robotics and Autonomous Systems Center (RASC), co-director of the USC Robotics Research Lab and Vice Dean for Research in the USC Viterbi School of Engineering. She received her PhD in Computer Science and Artificial Intelligence from MIT in 1994, MS in Computer Science from MIT in 1990, and BS in Computer Science from the University of Kansas in 1987.

Getting to Know Maja Matarić

How did you become interested in robotics and AI?

When I moved to the US in my teens, my uncle wisely advised me that "computers are the future" and that I should study computer science. But I was always interested in human behavior. So AI was the natural combination of the two, but I really wanted to see behavior in the real world, and that is what robotics is about. Now that is especially interesting as we can study the interaction between people and robots, my area of research focus.

Do you have any suggestions for people interested in doing outreach to K-12 students or the general public?

Getting involved with K-12 students in incredibly rewarding! I do a huge amount of K-12 outreach, including students, teachers, and families. I find the best way to do so is by including my PhD students and undergraduates, who are naturally more relatable to the K-12 students: I always have them say what "grade" they are in and how much more fun "school" is once they get to do research. The other key parts to outreach include letting the audience do more than observe: the audience should get involved, touch, and ask questions. And finally, the audience should get to take something home, such as concrete links to more information and accessible and affordable activities so the outreach experience is not just a one-off. Above all, I think it's critical to convey that STEM is changing on almost a daily basis, that everyone can do it, and that whoever gets into it can shape its future and with it, the future of society.

How do you think robotics or AI researchers in academia should best connect to industry?

Recently connections to industry have become especially pressing in robotics, which

has gone, during my career so far, from being a small area of specialization to being a massive and booming area of employment opportunity and huge technology leaps. This means undergraduate and graduate students need to be trained in latest and most relevant skills and methods, and all students need to be inspired and empowered to pursue skills and careers in these areas, not just those who self-select as their most obvious path; we have to proactively work on diversity and inclusion as these are clearly articulated needs by industry. There are great models of companies that have strong outreach to researchers, such as Microsoft and Google to name two, both holding annual faculty research summits and having grant opportunities for faculty to connect with their research and business units. As in all contexts, it is best to develop personal relationships with contacts at relevant companies, as they tend to lead to most meaningful collaborations.

What was your most difficult professional decision and why?

It's hard to pick one, but here are, briefly, three that are interesting: 1) I had to actively choose whether to speak up against unfair treatment when I was still pre-tenure and in a very underrepreresented group, or to stay silent and not make waves. I spoke up and never regretted being true to myself. 2) I had to choose whether to take part of my time away from research to get involved and stay involved in academic administration. I chose to do so, but also chose to never let it take more than the official half time, and never stomp on my research. 3) I had to choose whether to leave academia for a startup or industry. These days, that is an increasingly complex choice, but as long as academia allows us to explore and experiment, it will remain the best choice.

What professional achievement are you most proud of?

The successes of my students and of my research field. Seeing my PhD students receive presidential awards while having balanced lives with families and still responding to my emails just makes me beam with pride. Pioneering a field, socially assistive robotics, that focuses on helping users with special needs, from those with autism to those with Alzheimer's, to reach their potential. Seeing that field become established and grow from the enthusiasm of wonderful students and young researchers is an unparalleled source of professional satisfaction.

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

Nobody, no matter how senior or famous, knows how things are going to work out and how much another person can achieve. So when receiving advice, believe encouragement and utterly ignore discouragement. I am fortunate to be very stubborn by nature, but it was still a hard lesson and I see too many young people taking advice too seriously; it's good to get advice but take it with a grain of salt: keep pushing for what you enjoy and believe in, even if it makes some waves and raises some eyebrows.

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

I think about that when I talk to K-12 students; I try to tell them that it is fine to have a meandering path. I finally understand that what really fascinates me is people and what makes us tick. I could have studied that from various perspectives, including medicine, psychology, neuroscience, anthropology, economics, history... but since I was advised (by my uncle, see above) to go into computer science, I found a way to connect those paths. It's almost arbitrary but it turned out to be lucky, as I love what I do.

What is a "typical" day like for you?

I have no typical day, they are all crazy in enjoyable ways. I prefer to spend my time in face-to-face interactions with people, and there are so many to collaborate with, from PhD students and undergraduate students, to research colleagues, to dean's office colleagues, to neighbors on my floor and around my lab, to K-12 students we host. It's all about people. And sure, there is a lot of on-line work, too, too much of it given how much less satisfying it is compared to human-human interactions, but we have to read, review, evaluate,

recommend, rank, approve, certify, link, purchase, pay, etc.

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

Since I got involved with socially assistive robotics, I truly love all my research projects: we are working with children with autism, with reducing pain in hospital patients, and addressing anxiety, loneliness and isolation in the elderly. I share with my students the curiosity to try new things and enjoy the opportunity to do so collaborative and often in a very interdisciplinary way, so there is never a shortage of new things to discover, learn, and overcome, and, hopefully, to do some good.

How do you balance being involved in so many different aspects of the robotics and AI communities?

With daily difficult choices: it's an hourly struggle to focus on what is most important, set the rest aside, and then get back to enough of it but not all of it and, above all, to know what is in what category. I find that my family provides an anchoring balance that helps greatly with prioritizing.

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

*Wall*E*: it's a wonderfully human (vulnerable, caring, empathetic, idealistic) portrayal of a robot, one that has all the best of our qualities and none of the worst. After that, *Robot and Frank* and *Big Hero 6*.

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

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

AI Buzzwords Explained: Multi-Agent Path Finding (MAPF)

Hang Ma (University of Southern California; hangma@usc.edu)
Sven Koenig (University of Southern California; skoenig@usc.edu)
DOI: 10.1145/3137574.3137579

Kiva Systems was founded in 2003 to develop robot technology that automates the fetching of goods in order-fulfillment centers. It was acquired by Amazon in 2012 and changed its name to Amazon Robotics in 2014. Amazon order-fulfillment centers have inventory stations on the perimeter of the warehouse and storage locations in its center, see Figure 1. Each storage location can store one inventory pod. Each inventory pod holds one or more kinds of goods. A large number of warehouse robots operate autonomously in the warehouse. Each warehouse robot is able to pick up, carry and put down one inventory pod at a time. The warehouse robots move inventory pods from their storage locations to the inventory stations where the needed goods are removed from the inventory pods (to be boxed and eventually shipped to customers) and then back to the same or different empty storage locations (Wurman, D'Andrea, & Mountz, 2008).[1]

These order-fulfillment centers raise a number of interesting optimization problems, such as which paths the robots should take and at which storage locations inventory pods should be stored. Path planning, for example, is tricky since most warehouse space is used for storage locations, resulting in narrow corridors where robots that carry inventory pods cannot pass each other. Warehouse robots operate all day long but a simplified one-shot version of the path-planning problem is the multi-agent path-finding (MAPF) problem, which can be described as follows: On math paper, some cells are blocked. The blocked cells and the current cells of n robots are known. A different unblocked cell is assigned to each of the n robots as its goal cell. The problem is to move the robots from their current cells to their goal cells in discrete time steps and let them wait there. The optimization objective is to minimize the makespan, that is, the number of time steps until all robots are at their goal cells. During each time step, each robot can move from its current cell to its current cell (that is, wait in its current cell) or to an unblocked neighboring cell in one of the four main compass directions. Robots are not allowed to collide. Two robots collide if and only if, during the same time step, they both move to the same cell or both move to the current cell of the other robot. Figure 2 shows an example, where the red and blue robots have to move to the red and blue goal cells, respectively.

There are also versions of the multi-agent path-finding problem with different optimization objectives than makespan (such as the sum of the time steps of each robot until it is at its goal cell) or slightly different collision or movement rules. For example, solving the eight-puzzle (a toy with eight square tiles in a three by three frame, see Figure 3) is a version of the multi-agent path-finding problem where the tiles are the robots.

Researchers in theoretical computer science, artificial intelligence and robotics have studied multi-agent path finding under slightly different names. They have developed fast (polynomial-time) algorithms that find solutions for different classes of multi-agent path-finding instances (for example, those with at least two unblocked cells not occupied by robots) although not necessarily with good makespans. They have also characterized the complexity of finding optimal (or bounded-suboptimal) solutions and developed algorithms that find them. A bounded-suboptimal solution is one whose makespan is at most a given percentage larger than optimal.

Interestingly, it is slow (NP-hard) to find optimal solutions (Yu & LaValle, 2013c; Ma, Tovey, Sharon, Kumar, & Koenig, 2016), although a slight modification of the multi-agent path-finding problem can be solved in polynomial time with flow algorithms, namely where n un-

[1] See the following YouTube video: https://www.youtube.com/watch?v=6KRjuuEVEZs

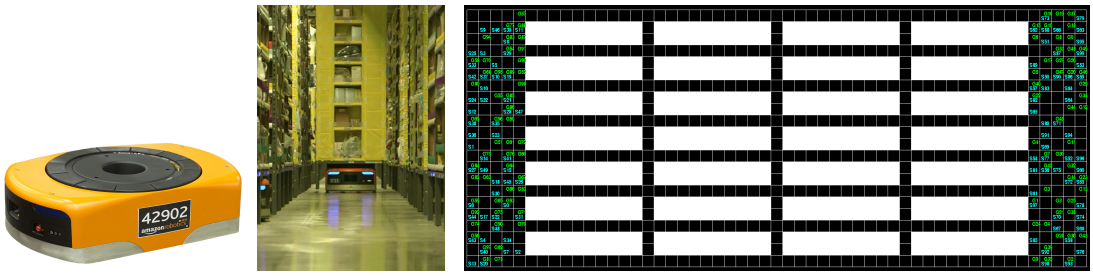


Figure 1: Warehouse robot (left), inventory pods (center), and the layout of a small simulated warehouse (right). The left and center photos are courtesy of Amazon Robotics.

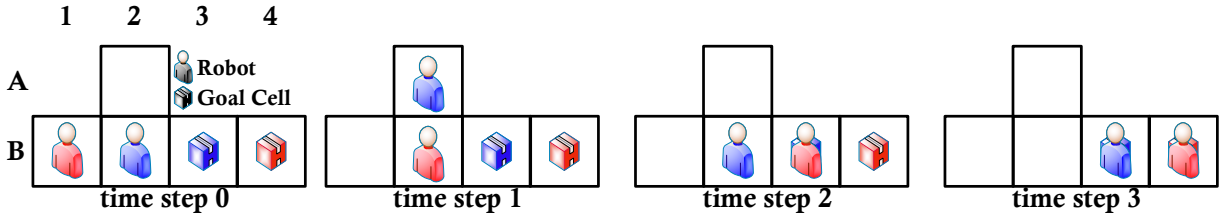


Figure 2: A multi-agent path-finding instance with two robots.

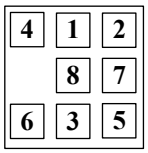


Figure 3: The eight-puzzle.

blocked cells are given as goal cells but it is up to the algorithm to assign a different goal cell to each one of the n robots (Yu & LaValle, 2013a). Researchers have also studied versions of the multi-agent path-finding problem where goal cells require robots with certain capabilities (Ma & Koenig, 2016) or robots can exchange their payloads (Ma et al., 2016).

In principle, one can model the original multi-agent path-finding problem as a shortest-path problem on a graph whose vertices correspond to tuples of cells, namely one for each robot, as shown in Figure 4 (where the red path shows the optimal solution), but the number of vertices can be exponential in the number of robots and the shortest path thus cannot be found quickly. Instead, researchers have suggested to plan a shortest path for each robot independently (by ignoring the other robots), which can be done fast. If all robots can follow their paths without colliding, then an optimal solution has been found. If not, then ...

- there are multi-agent path-finding algorithms that group all colliding robots together and find a solution for the group with minimal makespan (by ignoring the other robots), and then repeat the process. The hope is to find a solution before all robots have been grouped together into one big group (Standley, 2010; Standley & Korf, 2011).
- there are other multi-agent path-finding algorithms that pick a collision between two robots (for example, robots A and B both move to cell x at time step t) and then consider recursively two cases, namely one where robot A is not allowed to move to cell x at time step t and one where robot B is not allowed to move to cell x at time step t. The hope is to find a solution before all possible constraints have been imposed (Sharon, Stern, Felner, & Sturtevant, 2015).

These state-of-the-art multi-agent path-finding algorithms are currently not quite able to find bounded-suboptimal solutions for 100 robots in small warehouses in real-time. The tighter the space, the longer the runtime. Researchers have also suggested a variety of other multi-agent path-finding techniques (Silver, 2005; Sturtevant & Buro, 2006;

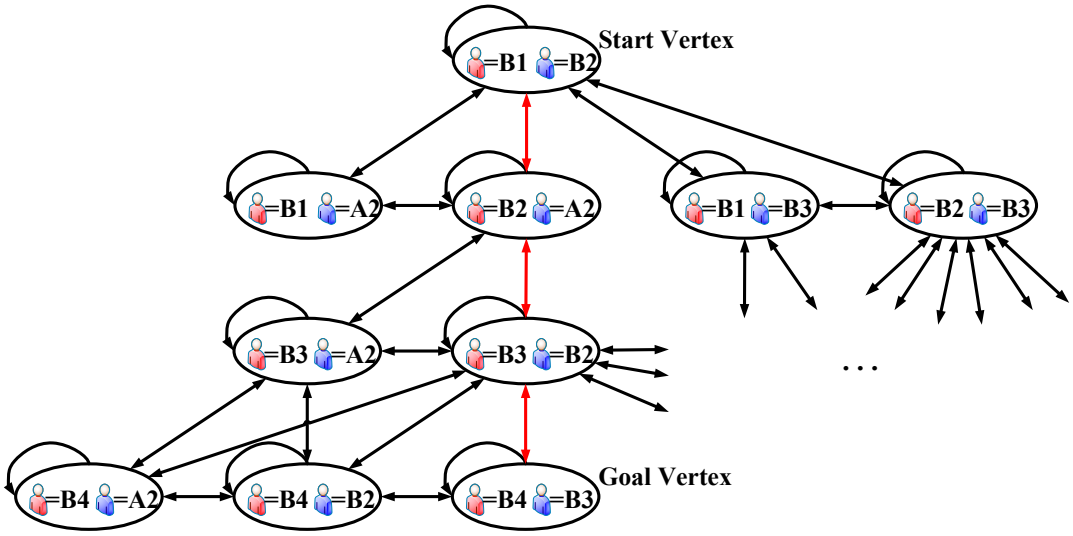


Figure 4: Partial graph for the multi-agent path-finding instance from Figure 2.

Ryan, 2008; Wang & Botea, 2008, 2011; Luna & Bekris, 2011; Sharon, Stern, Goldenberg, & Felner, 2013; de Wilde, ter Mors, & Witteveen, 2013; Barer, Sharon, Stern, & Felner, 2014; Goldenberg et al., 2014; Wagner & Choset, 2015; Boyarski et al., 2015; Ma & Koenig, 2016; Cohen et al., 2016), including some that transform the problem into a different problem for which good solvers exist, such as satisfiability (Surynek, 2015), integer linear programming (Yu & LaValle, 2013b) and answer set programming (Erdem, Kisa, Oztok, & Schueller, 2013). Researchers have also studied how to execute the resulting solutions on actual robots (Cirillo, Pecora, Andreasson, Uras, & Koenig, 2014; Hoenig et al., 2016).

Two workshops have recently been held on the topic, namely the AAAI 2012 Workshop on Multi-Agent Pathfinding[2] and the IJCAI 2016 Workshop on Multi-Agent Path Finding[3]. Recent dissertations include (Wang, 2012; Wagner, 2015; Sharon, 2016).

References

Barer, M., Sharon, G., Stern, R., & Felner, A. (2014). Suboptimal variants of the conflict-based search algorithm for the multi-agent pathfinding problem. In *Annual symposium on combinatorial search* (pp. 19–27).

Boyarski, E., Felner, A., Stern, R., Sharon, G., Tolpin, D., Betzalel, O., & Shimony, S. (2015). ICBS: Improved conflict-based search algorithm for multi-agent pathfinding. In *International joint conference on artificial intelligence* (pp. 740–746).

Cirillo, M., Pecora, F., Andreasson, H., Uras, T., & Koenig, S. (2014). Integrated motion planning and coordination for industrial vehicles. In *International conference on automated planning and scheduling* (pp. 463–471).

Cohen, L., Uras, T., Kumar, T. K. S., Xu, H., Ayanian, N., & Koenig, S. (2016). Improved solvers for bounded-suboptimal multi-agent path finding. In *International joint conference on artificial intelligence* (pp. 3067–3074).

de Wilde, B., ter Mors, A., & Witteveen, C. (2013). Push and rotate: Cooperative multi-agent path planning. In *International conference on autonomous agents and multi-agent systems* (pp. 87–94).

Erdem, E., Kisa, D., Oztok, U., & Schueller, P. (2013). A general formal framework for pathfinding problems with mul-

[2] See the following URL: http://movingai.com/mapf

[3] See the following URL: http://www.andrew.cmu.edu/user/gswagner/workshop/ijcai_2016_multirobot_path_finding.html

tiple agents. In *AAAI conference on artificial intelligence* (pp. 290–296).

Goldenberg, M., Felner, A., Stern, R., Sharon, G., Sturtevant, N., Holte, R., & Schaeffer, J. (2014). Enhanced partial expansion A*. *Journal of Artificial Intelligence Research*, *50*, 141–187.

Hoenig, W., Kumar, S., Cohen, L., Ma, H., Xu, H., Ayanian, N., & Koenig, S. (2016). Multi-agent path finding with kinematic constraints. In *International conference on automated planning and scheduling* (pp. 477–485).

Luna, R., & Bekris, K. (2011). Push and swap: Fast cooperative path-finding with completeness guarantees. In *International joint conference on artificial intelligence* (pp. 294–300).

Ma, H., & Koenig, S. (2016). Optimal target assignment and path finding for teams of agents. In *International conference on autonomous agents and multiagent systems* (pp. 1144–1152).

Ma, H., Tovey, C., Sharon, G., Kumar, T. K. S., & Koenig, S. (2016). Multi-agent path finding with payload transfers and the package-exchange robot-routing problem. In *AAAI conference on artificial intelligence* (pp. 3166–3173).

Ryan, M. (2008). Exploiting subgraph structure in multi-robot path planning. *Journal of Artificial Intelligence Research*, *31*, 497–542.

Sharon, G. (2016). *Novel search techniques for path finding in complex environment* (Unpublished doctoral dissertation). Ben-Gurion University of the Negev.

Sharon, G., Stern, R., Felner, A., & Sturtevant, N. (2015). Conflict-based search for optimal multi-agent pathfinding. *Artificial Intelligence*, *219*, 40–66.

Sharon, G., Stern, R., Goldenberg, M., & Felner, A. (2013). The increasing cost tree search for optimal multi-agent pathfinding. *Artificial Intelligence*, *195*, 470-495.

Silver, D. (2005). Cooperative pathfinding. In *Artificial intelligence and interactive digital entertainment* (pp. 117–122).

Standley, T. (2010). Finding optimal solutions to cooperative pathfinding problems. In *AAAI conference on artificial intelligence* (pp. 173–178).

Standley, T., & Korf, R. (2011). Complete algorithms for cooperative pathfinding problems. In *International joint conference on artificial intelligence* (pp. 668–673).

Sturtevant, N., & Buro, M. (2006). Improving collaborative pathfinding using map abstraction. In *Artificial intelligence and interactive digital entertainment* (pp. 80–85).

Surynek, P. (2015). Reduced time-expansion graphs and goal decomposition for solving cooperative path finding sub-optimally. In *International joint conference on artificial intelligence* (pp. 1916–1922).

Wagner, G. (2015). *Subdimensional expansion: A framework for computationally tractable multirobot path planning* (Unpublished doctoral dissertation). Carnegie Mellon University.

Wagner, G., & Choset, H. (2015). Subdimensional expansion for multirobot path planning. *Artificial Intelligence*, *219*, 1–24.

Wang, K. (2012). *Scalable cooperative multi-agent pathfinding with tractability and completeness guarantees* (Unpublished doctoral dissertation). Australian National University.

Wang, K., & Botea, A. (2008). Fast and memory-efficient multi-agent pathfinding. In *International conference on automated planning and scheduling* (pp. 380–387).

Wang, K., & Botea, A. (2011). MAPP: a scalable multi-agent path planning algorithm with tractability and completeness guarantees. *Journal of Artificial Intelligence Research*, *42*, 55–90.

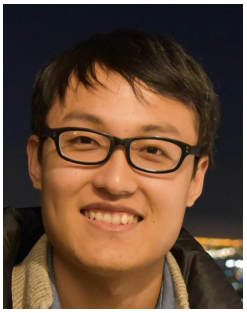
Wurman, P., D'Andrea, R., & Mountz, M. (2008). Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI Magazine*, *29*(1), 9–20.

Yu, J., & LaValle, S. (2013a). Multi-agent path planning and network flow. In E. Frazzoli, T. Lozano-Perez, N. Roy, & D. Rus (Eds.), *Algorithmic foundations of robotics x, springer tracts in advanced robotics* (Vol. 86, pp. 157–173). Springer.

Yu, J., & LaValle, S. (2013b). Planning optimal paths for multiple robots on graphs. In *Ieee international conference on robotics and automation* (pp. 3612–3617).

Yu, J., & LaValle, S. (2013c). Structure and intractability of optimal multi-robot path

planning on graphs. In *AAAI conference on artificial intelligence* (pp. 1444–1449).



Hang Ma is a Ph.D. student in computer science at the University of Southern California. He is the recipient of a USC Annenberg Graduate Fellowship. His research interests include artificial intelligence, machine learning and robotics. Additional information about Hang can be found on his webpages: http://www-scf.usc.edu/~hangma. Contact him at hangma@usc.edu.

Sven Koenig is a professor in computer science at the University of Southern California. Most of his research centers around techniques for decision making (planning and learning) that enable single situated agents (such as robots) and teams of agents to act intelligently in their environments and exhibit goal-directed behavior in real-time. Additional information about Sven can be found on his webpages: http://idm-lab.org. Contact him at skoenig@usc.edu.

AI Education: Deep Neural Network Learning Resources

Todd W. Neller (Gettysburg College; tneller@gettysburg.edu)
DOI: 10.1145/3137574.3137580

Introduction

In this column, we focus on resources for learning and teaching deep neural network learning. Many exciting advances have been made in this area of late, and so many resources have become available online that the flood of relevant concepts and techniques can be overwhelming. Here, we hope to provide a sampling of high-quality resources to guide the newcomer into this booming field.

Textbooks and Papers

Deep Learning (Goodfellow, Bengio, & Courville, 2016) is a popular recent textbook that seeks to briefly introduce background mathematical topics (e.g. linear algebra, probability and information theory, numerical computation) as well as machine learning basics. The second part of *Deep Learning* treats core material of deep learning practice (e.g. deep feedforward networks, regularization, convolutional networks, recurrent networks, etc.), whereas the third part covers topics of modern research interest in deep learning.

This textbook is available in HTML form on the authors' Deep Learning Book website and it is not difficult to find other e-book formats online that have been built from these HTML pages. However, this text alone is not the easiest introduction to the field. We would recommend Andrew Ng's Machine Learning MOOC, *An Introduction to Statistical Learning with Applications in R* (James, Witten, Hastie, & Tibshirani, 2014), and other resources listed in this AI Education Matters column in Volume 3, Number 2 as starting points for background material relevant to all machine learning.

Also recommended as a gentler introduction is Michael Nielson's *Neural Networks and Deep Learning* online book.

Web Resources

One of the best news feeds for following Deep Learning research developments and learning resources is Waikit Lau and Arthur Chan's *Artificial Intelligence and Deep Learning (AIDL)* Facebook group. At the time of this writing, it has over 30,000 members and features an active and steady flow of research results, tutorials, announcements, and Q&A discussion relevant to deep learning. Recommendations much like these can be found in questions 2-4 of the AIDL FAQ.

Other good sites for suggested starting points for learning about DL is *A Guide to Deep Learning* by YerevaNN Labs, Piotr Migdał's *Learning Deep Learning with Keras*, a16z team's reference links, Stanford's CS 231n *Convolutional Networks* course website, and, of course, various Wikipedia pages concerning artificial neural networks.

MOOCs

In April 2017, David Venturi collected an impressive list of Deep Learning online courses along with ratings data. In August 2016, Arthur Chan listed his top 5 lists. Concurring with these bloggers, we found Geoffrey Hinton's *Neural Networks for Machine Learning* course lectures to be a good high-level introduction to the field. However, this course is not oriented towards the beginner.

We recommend taking Andrew Ng's Machine Learning MOOC for background coverage and then supplementing Hinton's MOOC with applied tutorial exercises found elsewhere.[1] Kaggle is a data science competition website that features tutorials, datasets, and challenges that offer practical experiential learning opportunities. Beyond Hinton's general introduction, Arthur Chan also recommends Hugo Larochelle's graduate-level online Neural Network course.

[1] At time of writing, Andrew Ng has announced a new Coursera specialization in deep learning.

Software

There are several popular software frameworks that facilitate rapid prototyping of deep learning systems. Most popular is Google's TensorFlow. An even higher-level layer that has become popular is Keras, which can run as a layer on top of TensorFlow, Microsoft Cognitive Toolkit (CNTK), or Theano. To grasp how high-level Keras is, consider these small MNIST digit recognition training examples featuring multi-class logistic regression, single-hidden-layer neural network training, and convolutional network training implemented in under ten lines each.

Other popular software for deep learning includes Torch (and PyTorch), Caffe, MXNet, and DeepLearning4J. While Python appears to be the most popular language for deep learning development, support exists for other programming languages.

Hardware

Researchers seem to obtain hardware with GPUs that support fast deep learning in three main ways. One expensive route is to buy machines marketed directly for deep learning that typically have high-end GPU specifications and come preinstalled with popular deep learning software. However, there are numerous DIY tutorials for buying the necessary parts and putting together an inexpensive deep learning machine. These options are at the extremes of the high-cost/low-effort and low-cost/high-effort spectrum.

We would recommend a middle-ground approach for time-strapped faculty with tight budgets: Buy a high-end gaming machine (e.g. Dell's Alienware desktops) with good GPUs and install the necessary software as needed. Tim Dettmers has shared results of a recent GPU comparison study, and Ved's d4datascience blog entry describes the process of installing CUDA libraries and TensorFlow in detail.

Your Favorite Resources?

These are but a few good starting points for learning about deep neural network learning. If there are other resources you would recommend, we invite you to register with our wiki and add them to our collection at http://cs.gettysburg.edu/ai-matters/index.php/Resources.

References

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. (http://www.deeplearningbook.org)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: With applications in r*. Springer Publishing Company, Incorporated. (http://www-bcf.usc.edu/~gareth/ISL/)

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

Celebrating the Past, Present, and Future of Computing

Timothy E. Lee (Carnegie Mellon University; timothyelee@cmu.edu)
Justin Svegliato (University of Massachusetts Amherst; jsvegliato@cs.umass.edu)
DOI: 10.1145/3137574.3137581

Abstract

Timothy Lee and Justin Svegliato, two Student SIGAI Scholars, cover *The 50 Years of the ACM Turing Award Celebration*, which convened in San Francisco last June. The semicentennial celebration addressed the past, present, and future advancements of computing, ranging from deep learning and ethics to augmented reality and quantum computing.

As Student Scholars sponsored by SIGAI, we are grateful for the opportunity to be a part of *The 50 Years of the ACM Turing Award Celebration*. For two days in June, hundreds of professors, researchers, and students from across the globe gathered together in San Francisco to celebrate the legacy of the Turing Award (often referred to as the Nobel Prize of computing) and the incredible advances in computing over the last 50 years. The semicentennial celebration also honored this year's Turing Award recipient, Tim Berners-Lee, for inventing the World Wide Web and related networking technologies such as the Semantic Web.

After opening remarks from a Turing laureate each day, we heard from several panels that spanned the field of computing, ranging from deep learning and ethics to augmented reality and quantum computing. Every panel featured a distinguished moderator and several panelists that included Turing laureates, prominent researchers, and rising stars in the field. Interleaved with the panels were several short films. These films featured the life and work of the father of computer science, Alan Turing, and highlighted the Turing laureates' contributions, including those made to the field of artificial intelligence by the AI Turing Award laureates. We were honored by the attendance of several AI Turing Award recipients: Judea Pearl (2011 Turing laureate), Ed Feigenbaum (1994 Turing laureate), and Raj Reddy (1994 Turing laureate).

Figure 1: SIGAI at *The 50 Years of the ACM Turing Award Celebration*. Pictured from left to right: Timothy E. Lee, Yolanda Gil, and Justin Svegliato. The bronze bust of Alan Turing unveiled during the conference is also shown here.

The first panel *Advances in Deep Neural Networks* was particularly relevant to the SIGAI community. Moderated by Judea Pearl, the panel featured Michael Jordan (UC Berkeley), Fei-Fei Li (Stanford), Stuart Russell (UC Berkeley), Ilya Sutskever (OpenAI), and Raquel Urtasun (Toronto). As a popular area not only in AI but also in computing in general, deep learning has emerged as a powerful approach for enabling machine intelligence. Sutskever explained that neural networks are essentially tunable circuits that learn high-dimensional mappings from data. Deep neural networks have emerged from the confluence of several factors: the recent advances in hardware (the "oxygen" of neural networks according to Sutskever), the availability of massive datasets via the Internet, and the accelerated progress of data science.

Despite their promising results, a common theme emerged from the panel. Although effective in particular domains, deep learning in its current form cannot be the fundamental abstraction of machine intelligence sought after by researchers. There are many questions

about its long-term viability as the bedrock of machine intelligence. As Jordan argued, today's neural networks are deep architecturally, but not semantically. Pearl questioned whether these networks could reason about causality, a central theme in his foundational work on Bayesian networks.

Outlining the weaknesses of today's deep neural networks segued into a discussion on the types of intelligent behavior that humans exhibit but these networks currently lack, such as semantic understanding, contextual reasoning, abstraction, and reasoning under uncertainty, all of which are easily handled by humans despite little training data. Russell drew a fitting analogy between Allen Newell, Cliff Shaw, and Herbert Simon's General Problem Solver and the need for exponential computing with deep learning and the need for exponential *data*. In Russell's opinion, hoping to achieve "tabula rasa" machine intelligence with only deep learning may be infeasible in some—or all—domains due to the data demands, and so we must continue to search for better techniques. Li offered a similar anecdote from her work with ImageNet. With enough data, deep neural networks are by far the state of the art in object recognition, but they perform poorly and cannot reason effectively without massive datasets. In the case of robotics, Urtasun noted that being unable to model uncertainty well in deep learning is a considerable drawback in her current work on self-driving vehicles where algorithm robustness is critical.

Still, even with these shortcomings, deep learning performs quite impressively in narrow problems, such as computer vision, image captioning, and object segmentation. In some cases, such as AlphaGo, it enables decision-making capabilities that are superior to human intelligence. Several of the panelists agreed that deep learning has matured enough to be used in industry, but the search for machine intelligence must continue. Ultimately, the panel could be best summarized by Li's comments: we are entering the "end of the beginning" for AI. Deep neural networks may be one of our best existing tools for enabling the development of intelligent agents, but even greater breakthroughs are yet to come.

In addition to the deep learning panel, the opening day of the celebration also featured four other panels with many prominent researchers from industry and academia, along with a talk by 2008 Turing laureate Barbara Liskov that explored the history of computing. First, in *Restoring Personal Privacy without Compromising National Security*, Whitfield Diffie, 2015 Turing laureate, along with several leaders in security, cryptography, and networking discussed how governments could obtain useful information using backdoors and other intentional vulnerabilities to aid criminal investigations without jeopardizing the privacy of society. Following a short film on Alan Turing's life, we then turned to Vint Cerf, 2004 Turing laureate, and several other distinguished researchers in *Preserving Our Past for the Future*. They considered the problem of how to store data for centuries to come and whether corporations or governments should fund such an endeavor.

Later that day, *Moore's Law Is Really Dead: What's Next?* headlined the 1992 Turing laureate Butler Lampson. The panel explored the ways in which the field can continue the trend of exponential technological growth despite that Moores Law has continued to slow down. During the panel, a common theme emerged: researchers will eventually leverage special-purpose hardware and quantum computing to push the boundaries of computing forward.

At the end of the day, we heard from Raj Reddy in *Challenges in Ethics and Computing*. Given the increasing relevance of AI and machine learning, Reddy believes that ethical questions in computing have become more important than ever. Noel Sharkey added several important questions in light of recent progress in self-driving cars and machine learning. How can self-driving cars make decisions that were once reserved for humans in life and death situations? And how do we ensure that data-driven algorithms escape bias against minorities in the justice system?

On the final day of the conference, there were two panels on some of the most rapidly growing fields in computing following a talk from Donald Knuth, 1974 Turing laureate. *Quantum Computing: Far Away? Around the Corner? Or Maybe Both at the Same Time?* that featured the 2000 Turing laureate Andrew Yao

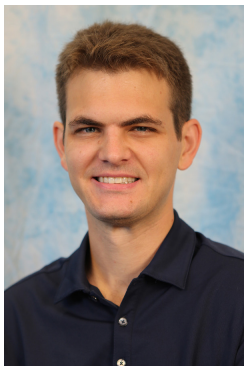
investigated the current state of quantum computing and how it might drive software development in the next 50 years. Like the deep learning panel, John Martinis cautioned that quantum computing is only a powerful tool in certain combinatorial problems but useless in others. However, in areas like AI, machine learning, and cryptography, it has the potential to revolutionize the field.

The celebration culminated with a panel on an area of computing that has recently seen rapid progress: *Augmented Reality: From Gaming to Cognitive Aids and Beyond*. Fred Books, 1999 Turing laureate, and Ivan Sutherland, 1988 Turing laureate, reminisced about the early work of augmented reality while they were aspiring researchers. Peter Lee discussed the impact of augmented reality on the gaming industry. Other panelists considered Google Glass, Pokemon Go, and Oculus Rift and explored the inevitable future of augmented reality in the home and at the workplace.

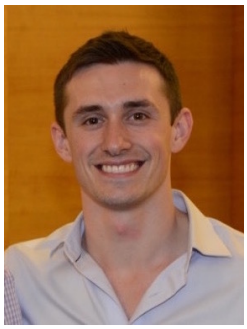
For all attendees across the spectrum of computing, the advancements of the last 50 years honored during the conference will undoubtedly shape our own contributions for the next 50 years to come. And, for the SIGAI community, the experts in our field gave insight into the ongoing search for machine intelligence and how deep learning might play a role. Given the recent groundbreaking advances in AI, it was only fitting that the celebration of computing's greatest achievements was in honor of who many call the grandfather of AI, Alan Turing.

Acknowledgements

We gratefully acknowledge the support of SIGAI for the opportunity to attend *The 50 Years of the ACM Turing Award Celebration*. We also thank Yolanda Gil for her help with this article.



Timothy E. Lee is a M.S. student in Robotics at Carnegie Mellon University. As a member of the Robust Adaptive Systems Lab under Professor Nathan Michael, Timothy's research focuses on improving mobile robot intelligence to enable the automation of challenging tasks in real-world settings. He is currently investigating robust, vision-based navigation of a submersible robot to automate the precision inspection of underwater infrastructure.



Justin Svegliato is a second year Ph.D. student in Computer Science at the University of Massachusetts Amherst. In the Resource-Bounded Reasoning Lab under Professor Shlomo Zilberstein, Justin's research focuses on bounded rationality, real-time decision making, and autonomous agent architectures. He is currently developing metareasoning techniques that monitor and control algorithms that trade decision quality with computation time.

ACM SIGAI CHINA:A New Incubator for AI in China

Le Dong (University of Electronic Science and Technology of China; ledong@uestc.edu.cn)
Man Yuan (University of Edinburgh; s1717356@sms.ed.ac.uk)
Ming-Liang Xu (Zhengzhou University; iexumingliang@zzu.edu.cn)
Ji Wan (Zhengzhou University; wanji@gs.zzu.edu.cn)
DOI: 10.1145/3137574.3137582

ACM SIGAI CHINA

The China Chapter of ACM SIGAI committees have been preparing ACM SIGAI China for a year. In March 2016, Le Dong and the ACM China council decided to establish ***a new chapter–ACM SIGAI China***.

In May 2016, the issue was further discussed in ACM China Council Meeting in Beijing. The Chair of ACM, Vicki L. Hanson, CEO, Robert B. Schnabel, and COO, Pat Ryan gave their approval to the foundation of ACM SIGAI China. The Chair of ACM China Council, Yunhao Liu from Tsinghua University and the Chair of ACM SIGAI, Sven Koening from University of California also gave their special support to the foundation of this new chapter in China. The issue was also reported to ACM and ACM China for permission. In October 2016, the organization structure of ACM SIGAI China was discussed in ACM China Council Meeting which was held in Taiyuan, China.

In March 2017, an official email from ACM informed that ***ACM SIGAI China chapter was successfully registered*** and ***ACM SIGAI China was officially founded***. In this ***new chapter, the Executive Chair & General Secretary is Le Dong*** from University of Electronic Science and Technology of China.

In this year, ACM SIGAI China symposium aims to provide ***a world's premier forum*** of renowned researchers to share their insightful opinions and discuss cutting-edge research on the artificial intelligence. The symposium features in types of sessions including distinguished talks and panel discussion. This summit forum expects to promote the development of artificial intelligence from the ***academic, technical to industry and applications***.

ACM TURC 2017 - SIGAI CHINA SYMPOSIUM

The ACM Turing 50th Celebration Conference - China (ACM TURC 2017), was held from May 12-14, 2017 in Shanghai, China with its key theme on "Trustworthy Network Big Data". The conference served as a highly selective and premier international forum on computer science research.

Figure 1: ACM TURC 2017

Alan Mathison Turing is the father of computer science, artificial intelligence, the founder of computer logic. He put forward the important concept of "Turing machine" and the "Turing test". In honor of this distinguished scientist, A.M. Turing Award is established according to his name in 1966 by the Association for Computing Machinery (ACM). It devotes to reward individuals who have made outstanding contributions to the computer industry. The Turing Award is recognized as the highest distinction in computer science and the Nobel Prize of computing.

On the occasion of the 50th anniversary of the establishment of A.M. Turing Award, ACM TURC 2017 was hosted by Shanghai Jiao Tong University, the Third Research Institute of the Ministry of Public Security, and Shanghai Municipal People's Government, co-organized by China Computer Federation, Tsinghua University, Peking University, Huazhong University of Science and Technology, and Tongji University.

In addition to the main sessions, the conference included ***workshops, panels, demonstrations***, and ***exhibits***. Twelve ***distinguished speakers*** including ***A.M. Turing Award Recipient*** were invited to discuss the frontier issues in today's society as well as the interdisciplinary development trends together.

ACM SIGAI is the Association for Computing Machinery's Special Interest Group on Artificial Intelligence. The Chair of ACM SIGAI is Sven Koening from University of California who is also AAAI Fellow and distinguished speaker of ACM.

This time, ***ACM SIGAI China*** invited many distinguished speakers to share their ideas with our members in the ***symposium***. They were Yangsheng Xu (Chinese University of Hong Kong), Sven Koenig (University of Southern California), Shipeng Li (CTO of Cogobuy Group and IngDan), Shiyi Chen (Fudan University), Bill Huang (CloudMinds Inc.), Gansha Wu (Uisee Technology Inc.), Gang Pan (Zhejiang University), Shuicheng YAN (Chief Scientist of Qihoo/360, Director of 360 AI Institute), Heng Tao SHEN (University of Electronic Science and Technology of China), Chunyuan Liao (Hiscene), Kai Yu (Horizon Robotics), Liang Lin (SenseTime Group Limited and Sun Yat-Sen University), Enhong Chen (University of Science and Technology of China), Junping Du(Beijing University of Posts and Telecommunications), Zenglin Xu (University of Electronic Science and Technology of China), Fei Wu (Zhejiang University), Rui Hou (Chinese Academy of Sciences), Zhongxing XUE (KingPoint), Jinglei ZHAO (ReadSense), Xuyao Hao, and Pianpian He(TF Securities). Yidan XU (Founder & CEO of Topplus), Linsen BAI (Founder of Alpha Brick), Wenzhi LIU (Engineering Director of SenseTime),and Yao Chen (Marketing Manager of Horizon Robotics) also delivered their speech in the Technical Review part. ***Our sponsors*** and ***enterprises*** also joined in the symposium panel discussion: ***From AI to Applications, what it takes to REALLY get there?***

In ***conference panel***, Vinton Cerf, Alexander Wolf, Wen Gao, Kai-Fu Lee, Xiangyan Li and Andrew discussed ***Big Data or Brain Powered Artificial Intelligence: Turing or Quantum?*** John Hopcroft, Recipient of the ACM

Figure 2: Le Dong & Kai-Fu Lee @ TURC 2017

Turing Award talked about ***Exciting ideas in computer science***. The ACM/IEEE Fellow, Gao Wen also brought up his ideas about ***Evolution of the Artificial Visual System***.

In the banquet, Haifeng Wang, Vice President of Baidu discussed ***AI Makes the Internet Smarter***. The Chairman and CEO of Sinovation Ventures, President of Sinovation Ventures Artificial Intelligence Institute and Science, Kai-Fu Lee also shared his ideas. Lee reviewed the history of computer and emphasized the application of quantum computer. He pointed out that today, people learn from mass data. ***In the future, people are able to learn from experience and have the ability of transfer learning***.

From the discussion, we can know that with the progress of science, AI will be used in ***cross-disciplinary programs*** and natural language understanding which will eventually be platformed and become the most effective tool which can be applied in deep learning, transfer learning, reinforcement learning, statistical learning, and so on. AI will eventually step into and have a huge impact on the physical world and also make remarkable accomplishments. But we have to be aware of the limitations of artificial intelligence. It lacks self-awareness, aesthetic, feelings or love. At the same time, it demands mass data, single domain and top scientists. People must be aware of these obstacles and realize that a new opportunity beckons in China. Human select AI as nature selects the fittest.There are several approaches for scientists: start their own business, empower technology, find a partner and publish papers.

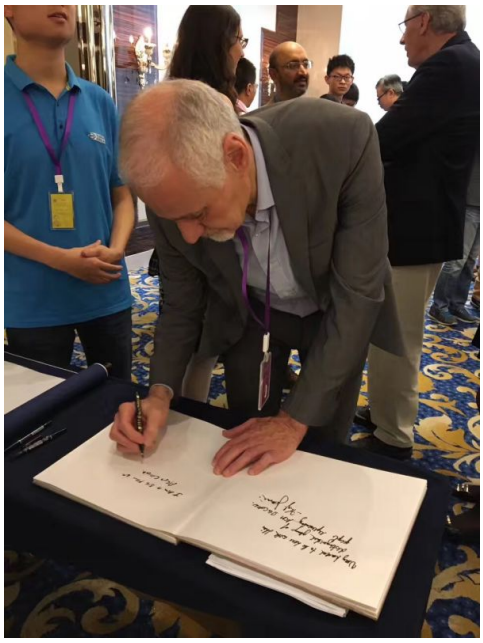


Figure 3: Alexander Wolf Left His Words in the Autograph Book @ TURC 2017

In the banquet, scientists, entrepreneurs and students exchanged their thoughts and ideas freely and left their precious signatures and some words in our autograph book.

The feature of Turing Conference is AI. In ***ACM SIGAI China Committee, Executive Chair & General Secretary is Le Dong*** from University of Electronic Science and Technology of China. Vice General Secretary are Mingliang Xu from Zhengzhou University and Yanli Ji from University of Electronic Science and Technology of China. Vice chair are Kun Zhou from Zhejiang University, Kai Yu from Horizon Robotics, Gansha Wu from Uisee Technology Inc., and Liang Lin from Sensetime Tech. Man Yuan and Ji Wan are also ACM SIGAI China secretary. In ACM TURC 2017, other organizers, XueLong Li, Hanli Wang, and Xiang Bai also made their contributions to this event. In particular, ***2017 Outstanding Contribution Award*** and ***Excellent ACM China Lecturer*** was presented to ***Le Dong*** by ACM Turing 50TH Celebration Conference - China for her extraordinary contributions to ACM TURC 2017 by ***Vicki Hanson, the Chair of ACM***.

Figure 4: Le Dong, The Executive Chair&General Secretary of ACM SIGAI CHINA Was Awarded by Vicki L.Hanson in The ACM Turing 50th Celebration Conference - China

The Future of AI

In May 4th, 2014, President Xi Jinping once pointed out that China is going to ***initiate world-class universities***.

In 2016, ***Report on the Work of the Government*** highlighted a new higher education policy, aiming to promote the construction of ***"Double First-Rate"***(developing First-class Universities and Academic Programs) in China. The development of cutting-edge technology relies heavily on higher learning institutions.

In 2017, President Xi attended the Opening Ceremony of ***Belt and Road*** Forum for International Cooperation(BRF) and he pointed out that in order to pursue innovation-driven development, China should strengthen the cooperation on the digital economy, artificial intelligence, nanotechnology, and quantum computing, and advance the development of big data, cloud computing and smart cities so as to turn them into a digital silk road of the 21st century. He also said that we should spur the full integration of science and technology into industries and finance, improve the environment for innovation and pool resources for innovation. Artificial Intelligence orients the development of future technology.

In July, 2017, ***the State Council*** notified ***the plan of the development for artificial intelligence*** in the following decades in China.

In such context, ***ACM SIGAI China*** is founded with three major missions:

First, ACM SIGAI China provides an ***incubator*** for this emerging industry. It connects AI and education, also with other fields. ***Academic faculty, student, enterprises*** are very welcomed to join ACM SIGAI China to promote the interaction and impact of disciplinaries in AI.

Second, we encourage ***original AI innovation*** and ***international cooperation*** within communities via various activities and events, through different branches in ACM and SIGAI, especially in China.

Third, we strongly advocate the variability of SIGAI CHINA and especially welcome ***students and females*** to become members of ACM SIGAI China.

We shall strengthen the context from senior researchers to junior students, as well as link partners among different regional areas. The new era of artificial intelligence needs every one of you! Learn and exchange the sparks of thoughts here!

More information about conference, recent events, registration, business cooperation and further details can be found on our website and WeChat Official Account. Your participation is warmly welcomed!

Figure 5: AI Helps Innovation

Find your opportunities at

Our website:

http://china.acm.org/TURC/2017/SIGAI.html

WeChat Official Account QR Code:

Le Dong received her PhD degree in Electronic Engineering and Computer Science from Queen Mary, University of London in 2009. She is a full professor in University of Electronic Science and Technology of China. She is the coordinator of several National Natural Science Foundation Projects such as NSFC Surface Project, NSFC Youth Project, NSFC Important Research Project and so on. She has now published more than 40 papers in international journals and conferences including several top journals and high level International Conference, such as TIP, ACM MM, TMM, PR, TCSVT, ICPR. She has served as a reviewer for several top journals and conferences. Now she is the Executive Chair & General Secretary of ACM SIGAI China , General Secretary of Vision And Learning Seminar (VALSE), General Chair of ACM TURC 2017 SIGAI CHINA Symposium, General Secretary of National Program of Global Experts, Youth Talents Committee, and the Executive Secretary of next-generation of Sichuan Provincial Engineering Laboratory. She got the Best Talk Award in the 10th ACM International Workshop on IOT and Cloud Computing, Excellent ACM CHINA Lecturer and ACM 2017 Outstanding Contribution Award in ACM Turing 50th Celebration Conference - China.

Man Yuan is currently a postgraduate student studying Teaching English to Speakers of Other Languages in Moray House School of Education in the College of Humanities and Social Science at the University of Edinburgh, Scotland, United Kingdom. She received her Bachelor of Arts degree in English from Shanghai International Studies University, and Bachelor of Science in Management Accounting as her minor specialty from Shanghai University of International Business and Economics, China in 2017. She has served as secretary of ACM SIGAI China, and translator in Huaqiao Foundation. Her research interests include English for specific purposes, teaching English to engineering and science students, language testing, and second language acquisition.

Ming-Liang Xu is a full professor in the School of Information Engineering of Zhengzhou University, China, and currently is the director of CIISR (Center for Interdisciplinary Information Science Research) and the Vice General Secretary of ACM SIGAI China. His research interests include computer graphics and artificial intelligence. He has now published more than 40 papers in international journals and conferences including ACM TOG, IEEE TPAMI, IEEE TIP, IEEE TCSVT. Xu got his Ph.D. degree in computer science and technology from the State Key Lab of CAD&CG at Zhejiang University.

Ji Wan is currently a postgraduate student majoring in Computer science and technology in Zhengzhou University. He received his Bachelor of Engineering degree in Computer science and technology from Zhengzhou University, China in 2015. He is the secretary of ACM SIGAI China. His research interests include mobile and spatial data management, location-based services , and smart city computing.

On the Importance of Monitoring and Directing Progress in AI

Lukas Prediger (RWTH Aachen University; lukas.prediger@rwth-aachen.de)
DOI: 10.1145/3137574.3137583

Abstract

This essay argues that the speed with which AI development proceeds will be an important factor for a beneficial adoption and thus should be closely monitored and controlled, if necessary. It also discusses privacy and manipulation, inequality of access and (value learning) superintelligence as major issues for all development trajectories.

Introduction

Recent years have seen steady progress in the development and application of artificial intelligence, mainly in the form of machine learning artifacts. The most prominent public achievements were DeepMind's AlphaGo mastering the Go board game last year (Hassabis, 2016) and, even more recently, Libratus, a program developed at Carnegie Mellon University, winning a match of Poker against professional players for the first time in history (Condliffe, 2017).

While these events mark important points in AI development, they represent only a small part of the "state of the art". The more important, but not as spectacular, development is the ever increasing number of narrow AI programs, especially the powerful combination of data mining and machine learning applications that help in recognizing patterns in all kinds of huge data bases, thus e.g. allowing companies to predict customer behavior. This greater ability to make predictions based on data is a major advantage and suggests that AI can yield massive benefits to whoever controls it as well as society as a whole. As an example, current research in almost all fields of science - especially natural and engineering sciences - relies on the ability to process data and was thus empowered and accelerated by computing machines. A greater ability to process data and (automatically) derive knowledge in the form of more accurate models and predictions as enabled by more sophisticated AI will consequently empower researchers further and likely help us solve problems that we struggle with today. More capable narrow AI can also reduce labor costs in manufacturing or services, yielding higher profit margins for companies and reduced consumer prices.

As a result, these AI applications receive tremendous interest in current research and can be expected to be further refined in the near future since they promise a very direct and obvious value to the companies applying them.

Other than AlphaGo and Libratus, these applications of AI already have a direct impact on society. For example, our notion of privacy is based upon the intuition that having more information about a person gives a greater ability to predict and manipulate this person's behavior in subtle or even open ways and thus access to this information should be restricted. Deriving information from seemingly innocouos data thus naturally raises concerns about the privacy of the data subjects. As companies gain more and more data on their customers, they can more accurately predict preferences and behaviors, not only to offer more specifically tailored products and services but also to influence customer decisions in ways that these might not even recognize.

One example for how such a manipulation might take place is the recent discussion on the extent to which content selection procedures in social networks might have skewed the public debate during the recent election in the United States.

Acknowledging this, it is clear that like any other technology AI opens up possibilities for benefits and harm alike and it is the responsibility of those who develop and employ it to take measures such that the benefits outweigh the possible harm. In parallel to the technical development, recent years have also seen a rising awareness about the ethical implications of AI, including (but not limited to) the possibility of massive job loss, privacy degradation, autonomous weapons, increases so-

cial inequity and more (cf. (Steinhardt, 2015; Brundage, 2015; Open Philantrophy Project, 2015)). There is also concern about possible superintelligent agents and a resulting existential threats to humanity (cf. (Bostrom, 2014)). The consensus is that AI will have a large - maybe unprecedented - transforming and probably irreversible impact on humanity. However, there is much uncertainty over how exactly this transformation will take shape, mostly because there is much uncertainty about the future progress in AI regarding both the levels of capability that can be achieved and how long it will take to reach each level. This translates into some uncertainty about the urgency and extent to which each of these concerns needs to be addressed.

I will argue in the first section that establishing a proper framework of AI development will be essential to streamline the discussion of measures to keep AI beneficial. However, there are also certain issues, namely its impact on the economy and labor market as well as privacy and manipulation hazards, that will almost certainly occur in the near future and need to be addressed immediately, as I will point out in the following sections, continued by a short argument about the long-term prospect of superintelligence. Note in advance that most the arguments I will state are not groundbreakingly new. I merely aim to emphasize their importance and add some minor points.

Monitoring and Predicting AI Development

Motivation

Much of the impact that AI will have depends on the speed of the development and application of new AI capabilities. Societies change constantly due to new circumstances, technologies or ideas but it usually takes several years, if not generations, for new paradigms to become accepted in mainstream opinion. This is especially true if they do not benefit everyone equally.

Disruptive technologies break up the prevalent composition of society and force it to adapt to new circumstances. Often this is due to a shift in employment because jobs are replaced by automation, pushing the workforce into sectors that cannot yet be automated. Not only does this change the job landscape but also the importance of certain skills and, by extent, the prestige of a person that holds them. Often, the values prevalent in a society change as behaviors enabled by new technology, while first usually regarded as strange and being rejected by the larger part of the population, become normal. However, as mentioned above, these processes usually take several years at least.

One of the reasons for this is that, since societies are complex systems, the consequences of a certain new technology or paradigm cannot be anticipated to a sufficiently accurate degree. Its implementation thus requires a monitoring during the process and adapting related regulatory measures in a reactive and iterative fashion as the impact becomes gradually more evident. This opens a window in which some aspects of new technologies might be unregulated and are open to exploitation.

Given these observations, the faster the disruptions caused by more advanced AI applications proceed and, consequently, the less time societies and regulatory bodies have to observe and adapt, the greater the probability and magnitude of societal instability or disorder is likely to be. Taking the labor market as an example, if jobs are automated in a gradual fashion, there is more time to retrain and educate those workers that lose their previous jobs. Further, assuming a gradual transition, there are more jobs still available for each wave of replaced workers and longer time windows for new jobs to emerge in the newly shaped economy. If, in contrast, waves of automation follow each other with very brief intermittent time periods, a large part of the workforce might be suddenly unemployed without having time to adapt during which the still employed support and smoothen the transition (cf. (Steinhardt, 2015)). Furthermore, while new job opportunities might open, they would probably also be swiftly automated without a sufficiently large part of the population having a chance to acquire the required skills to pursue them.

While these examples only focus on the labor market, the same arguments can be construed for other aspects of society in similar

ways.

Recent trends seem to suggest that the time for adoption of new technology into society is shortening (cf. (McGrath, 2013)), suggesting some kind of resilience of society to getting outpaced by technological advancement. However, the exact dynamics of this and whether it will hold up with more disruptive changes than simple convenience devices such as smartphones or whether there is some limit to the adoption speed (as suggested above) remain unclear. Research aimed in that direction, especially on variations between different subgroups of society, e.g. groups of age or ethnicity, might also reveal relevant measures to ensure a more stable and beneficial transition.

Modeling Progress

Following from the above, it seems paramount to have as accurate knowledge as possible about the speed with which AI development will probably progress and when to expect which changes in capability of these systems. Having this knowledge will allow us to anticipate the most disruptive changes in advance and smoothen their impact through preparatory measures or, if necessary, delaying development or implementation just enough to allow for a more gradual transition.

We are still in the dark about which exact qualities or properties make someone (or something) ”intelligent” and thus the final complexity of AI cannot be properly estimated. So far, we cannot even reliably establish how many scientific breakthroughs are approximately required or what the the ultimate final result of AI research will be. Without even knowing the required steps, it is certainly impossible to predict when they will occur. Establishing a precise model for future progress thus seems a to be a hopeless endeavor.

However, there might be ways to get a reasonably accurate understanding of the degree of capability by observing past trends and projecting them into the future. A prominent example of this is Kurzweil’s book (Kurzweil, 2005) in which he observes past trends in overall increasing complexity of biological and then technical systems and uses them to formulate a scenario of how future progress might play out. From this he derives concrete years when a certain new capability (e.g. whole brain emulation) is achieved, mostly from the ongoing exponential growth of computational power that he expects, combined with an estimation of the required computational power to achieve these feats. He also points out that even if his estimates are off by some orders of magnitude, this would only delay these technologies for a small number of years due to the exponential nature of advancing technology.

Critics pointed out that the exponential growth paradigm is not a reasonable assumption since past development of AI capability does not follow the increase in computational power in a linear fashion (cf. Myhrvold’s contribution to the Edge conversation (Brockman, 2014)). Recent findings however suggest that hardware development has a large contribution to current AI performance (Brundage, 2016) and that AI development seems to be, in fact, accelerating (Stone et al., 2016). It should nevertheless be pointed out that some of the milestones that Kurzweil predicts have already been missed, indicating that his estimations are, at least, overly optimistic. There was also criticism that there is no reason why events should play out in the order that he establishes.

However, there is some merit in the approach of construing one possible scenario and evaluating the impact is has, as pointed out by Goertzel (2007). Given the large uncertainty we are facing when predicting AI progress, it might be helpful to explore several scenarios in detail and refine them over time, as the trajectory of development we are really on becomes more clear.

As an example, a more recent effort by Stanford University’s AI100 project attempts to forecast how AI systems might be implemented in a typical American city in 2030, providing insight into medium-term development and pointing out possible concerns that should be addressed, as well as research directions to do so (Stone et al., 2016). Furthermore, the concrete issue of economic impact has already received significant interest and a wealth of literature exists (e.g. (Brynjolfsson & McAfee, 2014)). However, studies that try to estimate a more general impact of AI on society and additionally explore different sce-

narios and directions of possible development might yield significant additional and required insight.

Finally, any prediction into the future requires us to have a reasonable understanding of the current state of development. Unfortunately, attempts to accurately assess the past development are few and suffer from a vagueness surrounding the term of AI. Brundage reviews some recent attempts in (Brundage, 2016) and concludes that more research is required and proposes relevant directions that should be explored.

An Outline of Assumed Progress

Following the above proposal of constructing likely scenarios for future progress, I want to briefly establish what I believe will be a probable trajectory before addressing the issues resulting from it in the following section.

First, there seems to be no compelling argument as to why human-level intelligence should be theoretically unreachable in an AI implementation. There are several ways to achieve this as pointed out by Bostrom (2014), e.g. full brain emulation or mathematical-functional simulation of the brain or even a completely artificial solution. All of these either require or will reveal insights into how the human brain works. Human brain research and development of general AI are thus deeply intertwined endeavors and progress in both will advance at a similar rate.

Notwithstanding the current trend of narrow AI research, I do further believe that the incentives for implementing human-level AI are strong enough that it will be created at some point in the future (cf. (Brundage, 2015; Kurzweil, 2005)). Already there are companies supplied with large amounts of resources, such as Alphabet's DeepMind, working on this very task with impressive results such as the previously mentioned AlphaGo ((Silver et al., 2016)) as well as a system that learns to play classic Atari 2600 Games without any previous knowledge (Mnih et al., 2015). There seems to be a general trend to expand the narrow domains of AI systems as they grow more capable for their tailored tasks (e.g., autonomous cars have come a long way from simply driving in a desert to being able to navigate roads in traffic).

It seems uncertain whether or not qualitative superintelligence, i.e. AI that 'thinks' in ways superior to humans, is possible. However, any human-level AI could benefit from advancements in hardware technology which will with high probability enable it to gradually speed up the involved calculations and thus evolve into a 'speed superintelligence', i.e. an intelligence of similar capability as the human brain, but operating vastly faster (cf. (Bostrom, 2014)).

To me, the above appears to be a probable development and is kept very broad intentionally. I will not try to give any precise dates or milestones here. As pointed out above, this would need more rigorous thinking and research. However, the above outline will suffice for me to argue about some issues that are currently already present and might get magnified by the development I foresee.

Primary Concerns for Human Society

Privacy, Manipulation and Control Implications

As briefly mentioned before, the ability to extract knowledge about a person's beliefs, opinions and behavior not only allows to offer better services to that person but also makes him/her vulnerable to manipulation and exploitation. This is, in a sense, already happening and, so far, the protection of that privacy is lacking behind the ever new frontiers that might compromise it.

An infamous example of this was reported by Duhigg (2012). Allegedly, the discount store retailer Target used data it collected or bought about its consumers to predict the pregnancy of women. Given that the habits of new parents tend to break up, the intention was to exploit this knowledge to acquire new permanent customers by sending special advertisement and thus stimulating newly forming buying habits to incorporate shopping at Target stores. The article further mentions that customers did not initially respond well to the targeted advertisements, feeling unduly spied upon, to which target adapted by concealing the specialized pregnancy-related advertisements within unrelated product ads.

It should be pointed out that this article is not without criticism. Critics argue that the prediction algorithm was probably not as exact as

portrayed in the article, the misprediction rate was likely to be high and that ”Target mixes up its offers not because it would be weird to send an all-baby coupon-book to a woman who was pregnant but because the company knows that many of those coupon books will be sent to women who arent pregnant after all.” (Harford, 2014).

While this might be true, the underlying motivation of (subconsciously) influencing customers based on the knowledge extracted from data does not change even if the prediction is less accurate than described. Progress in AI will likely increase the reliability of these predictions and, in accordance to the intertwindness of AI and human brain research, offer new insights into how a person can be manipulated and directed. Left for exploitation, this is a scary prospect.

Of course, all of marketing and even most interactions between persons are manipulative to some extent and everyone counteracts such attempts on a daily basis. However, we usually do not have as large a difference in available data about the other party and as the possibilities to take influence grow, we need to explore the extent of manipulation that we deem acceptable.

To prevent these concerns from becoming real threats, governments and other regulatory bodies should follow closely on the developments and regulate the extent of the implementation of manipulating behavior. All the above issues also obviously apply to (state) surveillance agencies, leaving governments in a conflict of interest that I cannot see how to resolve for now.

A related issue is the willingness with which many software users surrender their private data in exchange for services. This is, to an extent, in itself an exploitation of the human risk-reward-system: The benefits offered by a service are most often obvious and immediate while the associated risks of giving up private data are abstract. Negative consequences might only occur after a long time, if at all, and might seem unrelated to the act of hading over the data. An emphasizing factor might be that users are dealing with systems instead of persons, which makes the surrendering of private data feel much more impersonal. It seems as if it is stored in some system for private future use. Overall, this leads to a devaluation and hence degradation of the notion of privacy in itself.

Left unchecked, this potential shift of value might continue to open up access to personal data until no person presides over his/her own personal data alone anymore, leaving that data, or at least parts of it, in the public domain. This must not necessarily happen and admittedly seems unlikely now, but the current trend hints in that direction. With it might come a greater favoring of interconnectedness and sharing between people instead of the individualism currently prevalent in Western societies.

Such a development must not be a bad thing but should not happen without reflection. People must be made aware of the data they provide and should be educated more thoroughly on the potential consequences. This should, however, not aim to evoke irrational fears causing rejection of technology but merely provide the necessary baseline for reflection and critical usage. If users know which information is required for a system to work, they are not only able to spot where unnecessary data is harvested but might also be able to increase the quality and thus usefulness of the required data they provide (cf. the Content Awareness privacy property described by Deng, Wuyts, Scandariato, Preneel, and Joosen (2011)).

An interesting topic of research to support or defeat the claim made above is whether people in Asian societies, where the collective and group is often valued higher than in Western ones, display a greater willingness to share (private) data.

Equality of Access

Another concern of importance is the equal access to AI systems and the required data to drive them. AI will empower whoever controls it by granting him/her/it superior information processing capability and automation of tasks, thus vastly increasing the overall capabilities of a single entity (person or institution). This bears the danger of opening a significant wealth-gap between those who have access to that technology and those who have not.

Access will most likely initially align itself with

the already existing rich-poor divide as the more wealthy can afford adopting new technology earlier. They might thus be able to establish a dominance in that field before the rest of the population is able to adopt, thus further improving their advantages following current trends (cf. (OECD, 2008; OECD, 2015)). Furthermore, at least in the early stages of AI development, greater knowledge of how the technology works will amplify its usefulness, putting higher educated persons at an advantage as well. Greater wealth typically also implies better education, suggesting a continued dominance of the currently wealthy. These are, of course, no new observations but might be amplified by emerging AIs of sufficient capability. We should ensure that every individual is sufficiently computer and data literate to be able to prevail in a society where AI based data processing is prevalent.

However, not only access to the hardware and algorithms that constitute an AI will be crucial, but also widespread availability of the relevant data. Without accurate and sufficiently large amounts of data, the usefulness of any AI will be severely restrained, putting again those who have access to that data at a significant advantage. Currently, the data collected by the large Internet companies (e.g. Facebook, Google, etc.) is and will likely continue to be the foundation of their business models, making open access to them unlikely. Policies that allow ordinary persons access to that data bases as part of a business plan (i.e. for a fee) can address this issue while covering the cost that these (and other) companies or maybe potential governmental agencies have in collecting and maintaining the data. Providing compensation for those who provide relevant data should be discussed (cf. (Brockman, 2014)).

The speed with which technology becomes available will again be a crucial factor in ensuring stability. Efforts at monitoring and, if necessary, establishing equality in access can more safely take place during a slow transition, making small adjustments along the way, than in a rapid progress that requires larger and more complex interventions with more uncertain outcomes at a larger scale.

Superintelligence

As pointed out before, I believe it is highly probable that superintelligent AI will be created at some point. There are serious concerns that this, if not handled correctly, might pose a serious threat to continued existence of human life (as we know and value it).

It is, of course, unreasonable to assume that an AI will arbitrarily decide to turn on its masters. We design these systems with the purpose to serve (some of) our goals encoded in them, otherwise there would be no incentive to create them. However, as pointed out by Omohundro (2014) and Bostrom (2014) and formalized by Benson-Tilsen and Soares (2016), there are certain instrumental goals such as self-preservation (including the protection of its current goal) and acquiring a maximal amount of resources that any AI agent striving to maximize any objective is extremely likely to converge to. Given that a sufficiently capable superintelligent agent is likely to outcompete ordinary humans and possibly human societies as a whole in any conceivable way, trying to control it by means of dominance is not a viable option. Hence, if a concern for human values is not sufficiently embedded in the agents overall goal, these instrumental goals will be a serious threat.

There is opposition to these concerns which mostly seems to take the point that it is either impossible or there is no incentive to create human-level and, consequently, superintelligent AI, that such a system would pose no threat, or that it is so far of in the future as to be irrelevant now (some of these arguments can be found in the Edge conversation (Brockman, 2014), stated in (Open Philantrophy Project, 2015) or in (Bryson, Kime, & Zürich, 2011)). However, given the arguments laid out so far, I do not find these counterclaims convincing.

It is thus important to solve this so-termed control problem before the advent of advanced AI. With the current high uncertainty in AI progress which makes it unforeseeable when this might occur and the additional uncertainty about the difficulty of solving the control problem, research on this should not be postponed.

Fortunately, recent years have seen an increasing interest towards this and started to

explore possible solutions. A currently often proposed method is the self-learning of human values from observing behavior and cultural evidence by the agent instead of trying to encode them by hand (cf. (Bostrom, 2014)). However, there are certain objections to this that should be carefully investigated.

Caliskan-Islam, Bryson and Narayanan (2016) observe that the semantics of natural language already encode bias and harmful prejudices and that systems learning associations from language corpora pick up these biases. This suggests that an agent that is supposed to learn human values from the evidence it sees will not only pick up those that we would label as good but also all the underlying biases we have towards each other that might harmfully skew the ultimate values it comes to extract. While embedded in society these prejudices are already harmful but we can examine them and try to find remedies. In a superintelligent agent they could be beyond correction and vastly more harmful.

Further, as pointed out by Isaksen, Togelius, Lantz, and Nealen (2016), humans engage in games of dominance with animals (i.e., beings of lesser intelligence) that sometimes involve even the death of the animal. This might suggest a tendency towards the exertion of dominance, including violence, if no significant resistance is to expect, that a value-learning AI might also pick up. Hobbes (1651) suggested that societies provide a stable and peaceful environment because of the limited ways in which single individuals can exert their powers over the many - mostly due to the fact that any single entity will be easily overpowered by a conglomerate of many entities if all have similar capability. If this were true, a superintelligent AI that embodies these human tendencies and is unconstrainable in this sense will be dangerous.

The ultimate point made here is that human behavior often does not align with the noble values we claim to hold. Ultimately, between different individuals, groups and cultures, we appear to not even agree on the ultimate values of humanity (although some baseline has been established). To ensure that value-learning will load an agent with truly benevolent goals we might need to find solutions to our conflicting values and misaligned behavior as displayed by humans first - before trying to create some uncontrollable entity that might incorporate them.

This last argument also further strengthens the necessity of equal access mentioned in the previous section. If only a reasonable balance of power between its constituents maintains the stability of a society and access to AI has the ability to massively empower individuals, ensuring equality of access is of tremendous importance.

Conclusion

In this essay I tried to lay out some of the aspects of continued development and integration of AI into human societies that I found most concerning and supply some arguments to the discussion. Since AI's impact will be enormous and is hard to anticipate exactly, it might be that my prioritization is wrong and other issues turn out to be more pressing. Military AI applications come to mind, for example. An ongoing discussion and future research of the possible and the monitoring of the real progress of AI development will yield insight into this over time.

The discussed issues have an immediate importance and will likely remain relevant over the entire time frame of transitioning into an AI empowered society (and possibly thereafter), so working on solutions to them will address long-term and short-term concerns in like manner.

I want to emphasize that, while it has serious concerns attached to it, AI also has an enormous beneficial potential in helping us solving outstanding problems and create sufficient wealth for everyone to profit from, potentially ending poverty on a global scale. By no means should these developments be suppressed out of misplaced fear. However, care should be taken that it is these beneficial consequences that come to pass, which should not be taken for granted if we are lacking the will or care to shape the development.

Most of these measures are more in the realm of policy making, either for governments or the research community, than in technical AI research and thus involve an open and honest discussion outside of a purely academic set-

ting. The overall public must have knowledge of and agency in the involved decisions. Given the global scale of AI impact, discussion and policy decisions must be globally coordinated. In potential impact, vagueness of threat (due to the long time scales) and difficulty to address, we are facing issues on a scale similar to climate change. That we have been unable to address the latter in a permanent way so far might well be a disheartening omen.

However, a recent surge in discussion of ethics and AI as well as AI impact displays that, at least in parts of the academic community, the issues have been realized and action is taking place. This can be seen in part by the many great works referred to in this text that provide amazing insights as well as the establishment of several institutes concerned with these topics or codes of ethics for AI research as well as research proposals. Given that, between some overly optimistic and pessimistic views, it seems that we are on an overall stable trajectory so far. If we do not stray from it and watch our steps closely, the future, in so far as it is affected by AI, might turn out fine.

References

Benson-Tilsen, T., & Soares, N. (2016). Formalizing convergent instrumental goals. In *2nd International Workshop on AI, ethics and society at AAAI-2016. Phoenix, AZ.* Retrieved from http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12634/12347

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. OUP Oxford.

Brockman, J. (2014, November 14). The myth of AI - A conversation with Jaron Lanier. *Edge*. Retrieved from https://www.edge.org/conversation/jaron_lanier-the-myth-of-ai

Brundage, M. (2015). Economic possibilities for our children: Artificial intelligence and the future of work, education, and leisure. In *1st International Workshop on AI and ethics at AAAI-2015.* Retrieved from http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10155

Brundage, M. (2016). Modeling progress in AI. In *2nd International Workshop on AI, ethics and society at AAAI-2016. Phoenix, AZ.* Retrieved from http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12662/12348

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies.* WW Norton & Company.

Bryson, J. J., Kime, P. P., & Zürich, C. (2011). Just an artifact: Why machines are perceived as moral agents. In *IJCAI proceedings - International joint conference on artificial intelligence* (Vol. 22, p. 1641).

Condliffe, J. (2017, January 31). An AI poker bot has whipped the pros. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/603544/an-ai-poker-bot-has-whipped-the-pros/

Deng, M., Wuyts, K., Scandariato, R., Preneel, B., & Joosen, W. (2011). A privacy threat analysis framework: Supporting the elicitation and fulfillment of privacy requirements. *Requirements Engineering*, *16*(1), 3–32.

Duhigg, C. (2012, February 16). How companies learn your secrets. *The New York Times*.

Goertzel, B. (2007). Human-Level artificial general intelligence and the possibility of a technological singularity: A reaction to Ray Kurzweil's The Singularity Is Near, and McDermott's critique of Kurzweil. *Artificial Intelligence*, *171*(18), 1161–1173.

Harford, T. (2014, March 28). Big data: Are we making a big mistake? *Financial Times*. Retrieved from https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0

Hassabis, D. (2016, March 16). What we learned in Seoul with AlphaGo. *Google's 'The Keyword' Blog*. Retrieved from https://blog.google/topics/machine-learning/what-we-learned-in-seoul-with-alphago/

Hobbes, T. (1651). *Leviathan*. London, Michael Oaskeshott edition.

Isaksen, A., Togelius, J., Lantz, F., & Nealen, A. (2016). Playing games across the superintelligence divide. In *2nd International Workshop on AI, ethics and society*

at AAAI-2016. Phoenix, AZ. Retrieved from http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12645/12350

Islam, A. C., Bryson, J. J., & Narayanan, A. (2016). Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, *abs/1608.07187*. Retrieved from http://arxiv.org/abs/1608.07187

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Penguin Books.

McGrath, R. (2013, November 25). The pace of technology adoption is speeding up. *Harvard Business Review*. Retrieved from https://hbr.org/2013/11/the-pace-of-technology-adoption-is-speeding-up

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

OECD. (2008). *Growing unequal?: Income distribution and poverty in OECD countries*. OECD Publishing. Retrieved from http://www.oecd-ilibrary.org/social-issues-migration-health/growing-unequal_9789264044197-en doi: 10.1787/9789264044197-en

OECD. (2015). *In it together: Why less inequality benefits all*. OECD Publishing. Retrieved from http://www.oecd-ilibrary.org/employment/in-it-together-why-less-inequality-benefits-all_9789264235120-en doi: 10.1787/9789264235120-en

Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 303–315.

Open Philantrophy Project. (2015, August). *Potential risks from advanced artificial intelligence.* Retrieved from http://www.openphilanthropy.org/research/cause-reports/ai-risk

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Steinhardt, J. (2015, June 24). Long-Term and short-term challenges to ensuring the safety of AI systems. *Personal Blog 'Academically Interesting'*. Retrieved from https://jsteinhardt.wordpress.com/2015/06/24/long-term-and-short-term-challenges-to-ensuring-the-safety-of-ai-systems/

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., ... Teller, A. (2016, September). *Artificial intelligence and life in 2030* (One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel). Stanford University, Stanford, CA. Retrieved from https://ai100.stanford.edu/2016-report

Lukas Prediger is a Master's degree student in Computer Science at RWTH Aachen University, Germany, with a recent stay at Aalto University, Finland, during which this essay came to be. His main study interests are artificial intelligence, machine learning, big data and IT security and all their entailed consequences for society, individual freedom and privacy.

Truth in the 'Killer Robots' Angle?

Matthew Rahtz (University of Zürich/ETH Zürich; mrahtz@ethz.ch)
DOI: 10.1145/3137574.3137584

I had no idea, getting interested in AI two years ago, that being involved in the field would involve such a persistent sense of unease. I started out unequivocally excited, perhaps a little naive; but over the years, the concerned voices of economists, philosophers, and the mass media have gradually seeped into me, leaving me with an ill-defined feeling of hesitation about what we're heading towards.

Trying to understand the situation a bit better over the past months has been somewhat overwhelming. AI touches so many different areas that it's hard to hold everything in mind at once. While there are points of concern in many of these areas, there are three areas in particular which have stood out to me: unemployment, the long-term risks of AI, and lethal autonomous weapons systems (otherwise known as 'killer robots').

Reading into these areas in more depth has left me surprised. AI risk research is something that's interested me for a while, but the pressing issue right now doesn't seem to be the immediate need for research - rather, the connotations that are becoming associated with that research. With the second of these areas, autonomous weapons, I began looking into it only for the sake of completeness, but found myself completely unaware of the gravity of the situation. The biggest surprise, however, has been a change of opinion about the danger of unemployment. I realise I am no longer nearly so concerned about the imminent threat of automation.

Given all that we've been hearing on the topic of AI-related unemployment recently, this last statement clearly requires some explanation. So that the scene is set properly for the former issues, let us in fact begin our discussion with this last point.

Unemployment

I was quite prepared to spend this essay arguing that unemployment resulting from use of AI was by far the most pressing issue right now. The more I've read, though, the more I've got the impression that the change is going to be slower than it might appear.

At the start of my reading, the threat of unemployment seemed clear. Automation in various forms has been encroaching on the job market for centuries. With the recent step change in our machine learning capabilities, it seemed a very reasonable concern that we may not be far away from a step change in automation, and hence unemployment, too.

Take autonomous vehicle technology. Companies like Google and Tesla seem to be getting pretty good at it. For most of us, self-driving cars are going to be an unambiguously good thing; promises of safer roads and more leisure time abound. The question is: what happens when that technology makes, say, self-driving trucks possible? There are 1.6 million long-haul truck drivers in the US (McArdle, 2015). It's the most popular job in 29 states (Solon, 2016). It's one of the last jobs left offering middle-class pay without a college degree (Kitroeff, 2016). When autonomous vehicle technology reaches the trucking industry and makes all these drivers redundant – what then?

Despite the apparently obvious problem, I was surprised to find that some people don't seem to be all that worried. The common arguments I've heard for this position, though, haven't convinced me.

"Jobs will be lost, yes; but the AI revolution will create new jobs at the same time," some say. But I see no guarantee that the jobs created will match the skills of the people being made redundant. This is already a problem: the global talent gap. The issue is not that there aren't enough jobs. The issue is that the unfilled jobs require skills that the unemployed population don't have.

Others say, "We've seen step changes in employment before; what about the industrial revolution? We survived that alright." But this doesn't comfort me, because the world is such a different place now. The internet, for example, enables advancements in technology to spread much quicker than ever before; and this is especially significant when the relevant technology is software. In general, it seems like a bad idea to try and make predictions based on only superficially-similar historic precedents.

The difficulty of making predictions in a highly unpredictable world is clearly one of the major factors limiting the quality of these discussions. Whatever ideas one may have, it's hard to avoid concluding with anything but the inevitable cop-out of, "But then again, technology changes so rapidly, who knows what may happen?"

This got me to thinking: what are the *technology-invariant* factors here? What dynamics will stay relevant regardless of what new kinds of technology we come up with? Of these factors, there's been one in particular that's struck me as significant: the Pareto principle. And there's no better example of this principle than in the development of self-driving cars.

My impression with autonomous vehicle technology is that we're a lot further away from complete human replacement than one might think at first glance. Sure, we've seen the exciting demos of self-driving technology. But while these demos are impressive, it's worth bearing in mind that even back in 2015, a single talented hacker could get similar demo-level functionality working in about a month (Vance, 2015). The difficulty is apparently not in getting something basically working; the difficulty is in getting it to work reliably, in a wide range of conditions. As Tesla point out in their response to said hacker's efforts: "This is the true problem of autonomy: getting a machine learning system to be 99% correct is relatively easy, but getting it to be 99.9999% correct, which is where it ultimately needs to be, is vastly more difficult." (Tesla, 2015)

This dynamic is, essentially, what the Pareto principle states: that the first 80% of the results tends to be achieved with the first 20% of the efforts (though the exact proportions don't matter). And it's this dynamic that makes me think that change is going to be much slower than we might expect.

Even Google's efforts seem to be in line with this principle. Their self-driving buggy without even a steering wheel is pretty cool, but there's a catch: it can only safely go 25 mph (Farivar, 2015). Also, it doesn't work in snow (McArdle, 2015). Turning back to self-driving trucks, if it's taken *Google* this long to get to achieve autonomy in even these limited conditions, I'd guess it's going to be a very long time before complete autonomy can be achieved for a multi-tonne truck travelling at high speed down a freeway in rain, sleet, fog and, indeed, often snow.

It's unsurprising, therefore, that current efforts at truck automation, such as those from Daimler (Davies, 2015) and recent startup Otto (Lee, 2016), are instead targeting semi-autonomous solutions. In good conditions, the truck will drive itself. In bad conditions, a human driver in the cab can take over. In good conditions, you still get the benefits of machine control, like the ability to drive throughout the night. But you're spared the difficulty of pushing all the way to the "99.9999%" that's required for complete automation.

I suspect this is a pattern we'll see throughout many industries. Sure, new technologies are going to pop up. And we'll see those technologies progress to the level of useful semi-automation pretty quickly. But it's going to take much longer for the technology to mature to the point where complete automation is possible.

This is not to say that complete automation won't happen eventually. But because of the Pareto principle, I think the change is going to be gradual. We're going to get advance warning of what's happening. It seems unlikely that it's going to be a step change.

I also don't mean to suggest that eventual wide-scale automation isn't something worth thinking about and preparing for. Indeed, I'm glad to see the issue receiving as much attention as it is. I only mean to say that it may be a better use of our energies right now to focus on other areas which are more pressing and, perhaps, more neglected.

This brings us to the first of what I believe

our current concerns really are: lethal autonomous weapons systems.

Lethal Autonomous Weapons Systems

One of the tropes brought up often in the media over the past few years has been the image of 'killer robots'. For a long time, the hyperbole that invariably accompanied such media lead me not to take the issue seriously. Even in retrospect, I'm not surprised at my ignorance. Those kinds of discussions never touched on what seems to be the *real* issue: the consequences of an autonomous weapons arms race.

First, let's clarify what we're talking about. Lethal autonomous weapons systems (LAWS), as they're known more dryly, refer to weapons which can make the decision to kill entirely on their own, without explicit go-ahead from a human. For example, the drones currently in use don't fall into this category, because the decision to kill must be made by a remote human operator. What we're talking about is, say, a drone that can find and kill a target while being completely disconnected from a pilot.

LAWS have already existed for a while – think land mines. More recently, though, advances in AI are starting to enable more sophisticated forms of LAWS. For example, automated sentry guns have already been developed and deployed along the border between North and South Korea (Rabiroff, 2010). This trend looks set to continue: with the advantages of LAWS (e.g. immunity to communications jamming; potentially more precise and accurate targeting; fewer soldiers' lives on the line), many nations are now investing heavily in their further development (Goose & Wareham, 2017). The question we now face is: should we allow this trend to continue?

There are arguments both ways. On the one hand, avoiding danger to soldiers' lives, LAWS lower the threshold of entry to conflict. And because of the difficulty in distinguishing combatants from non-combatants, LAWS could lead to an increase in the number of civilian casualties (Goose & Wareham, 2017). On the other hand, if we *can* program them with humanitarian law, perhaps LAWS could be more ethical than their stressed-out human counterparts (Arkin, 2015).

The most forceful argument I've come across, however, concerns the likely consequences of a LAWS arms race. The idea is: once one nation starts deploying sophisticated LAWS, other countries will feel the need to step up their own efforts to develop and deploy LAWS of their own, leading to a positive feedback loop (Future of Life Institute, 2015b). That race is going to lead to even more sophisticated forms of LAWS being developed, and at ever lower prices. With proliferation happening all around the world, at some point it seems inevitable that some units will fall (or be sold) into the wrong hands. Consider "the availability on the black market of mass quantities of low-cost, anti-personnel microrobots that can be deployed by one person to anonymously kill thousands or millions of people who meet the user's targeting criteria" (Russell, Tegmark, & Walsh, 2015), and you get the picture.

The proposal, therefore, is to ban LAWS before this arms race can get started.

Indeed, this is the direction that the gears of international government - the UN Convention on Certain Conventional Weapons (CCW) – *are* moving in. The problem is that they may not be moving fast enough. Only at the end of 2016, after three years of discussion, has the CCW agreed to establish a Group of Governmental Experts under whom the creation of new international law can be discussed (Wareham, 2017a). Whether this group will move quickly enough to prevent the start of the race is still uncertain (Wareham, 2017a). It's not even clear whether this discussion will really lead to a complete ban, or only regulation limiting LAWS' use (Goose & Wareham, 2017). Given that deployment of sophisticated LAWS may be only years away (Future of Life Institute, 2015b), we're at a decisive moment.

Reading about LAWS, I've been forced to admit that the 'killer robot' angle really does have a grain of truth in it. There is, however, a second angle of the scare I've gradually become convinced it's worth taking seriously: the long-term risks of AI. But it's not the risks themselves that I think are the most pressing issue right now. The bigger issue at the moment is the culture that's becoming associated

with such concerns – and the limiting effect that culture might have on the field's growth. This brings us to our second pressing issue: opinion of AI risk research.

Opinion of AI Risk Research

Though the dangers of 'killer robots' have been talked about for decades, research into the long-term risks of AI only seems to have started being taken seriously with the publication in 2014 of Nick Bostrom's book 'Superintelligence' (Bostrom, 2016). Bostrom argues that the real threat will come not from robots, but from artificial *general* intelligence (AGI): AI which is superhumanly capable across a wide range of different tasks, rather than just the narrow domains that current AI can deal with. Consider AI with superhuman cognitive abilities (without the rest of our ancestral baggage, like emotions) that can be put to work on arbitrary problems, and you get the idea.

Such technology is, of course, not around the corner. Reflecting, though, that over the course of my father's life, we went from complete ignorance of DNA to being able to precisely engineer super-muscular dogs (Regalado, 2015), and from complete lack of digital technology to small devices we can fit in our pockets with radio access to the sum of all human knowledge, it seems within the realms of possibility that AGI may happen within our lifetimes. AGI may be a way away, but not so far as to be completely intangible to us.

Despite being such a long way away, Superintelligence concludes, AI risk research is nonetheless something we need to start working on *now*. Why? Because the advent of AGI is likely to be one of the most momentous events in the history of mankind. After AGI, we're likely to be forced onto one of two paths. There's the 'bad' path, where, for example, AGI allows one organisation or state to assume control; or where an errant AGI set up with an faulty objective gradually consumes all the world's resources in order to achieve its goal. But there's also the 'good' path, where AGI allows us to solve a whole slew of problems that have thus far proved beyond the reach of our small, metabolically-limited brains. Given the all-or-nothing nature of the outcome, Superintelligence argues, and given that it may be difficult to alter our trajectory once popularity of potentially unsafe algorithms has passed some critical level, it's worth us getting started as early as possible on making sure that we get the *good* path.

The most pressing issue right now, however, isn't the immediate need for research. Yes, the proportion of the AI community that's dedicated to risk research is less than what it ideally would be. But given the long-term nature of the problem, what's important is not the starting level, but the rate of growth the field will experience over the coming decades. And this is where I get worried.

The publication of Superintelligence around the middle of 2014 succeeded in bringing awareness of the issue to a broader audience. Some of that audience were in a position to rebroadcast to a broader audience still: over the subsequent months, we saw public statements of concern from the likes of Elon Musk in October 2014 (Gibbs, 2014), Stephen Hawking in December 2014 (Cellan-Jones, 2014), and Bill Gates in January 2015 (Mack, 2015). Though beneficial in publicising the issue, taken out of the context of the broader discussion, these statements seem to have had some undesirable consequences.

One of the consequences has come about through the coincident media hype about recent advances in machine learning. This seems to have given some the impression that the concern about the risks of AGI is based on an assumption of imminence. A report in January to the US Department of Defense by the JASON advisory group, for example, states that "the claimed 'existential threats' posed by AI seem at best uninformed. . . in the midst of an AI revolution, there are no present signs of any corresponding revolution in AGI" (JASON, 2017). But this is not the basis for the concern at all. In fact, it's a measure of just how seriously those involved believe the future dangers to be that *even though* AGI is likely to be a very long way away, it's still worth preparing for now. It would be unfortunate if misunderstanding on this point were to lead to misallocation of resources.

There is, however, a second, more serious consequence. These statements, combined with the above-mentioned Terminator articles, seem to have created a media at-

mosphere where questions about the dangers of AI have become appealingly provocative (Bostrom, 2016). This provocation has, in turn, seem to have stirred up an (understandable) feeling of defensiveness among some in the AI community. Mustafa Suleyman, one of the co-founders of Google DeepMind, for example, was quoted at a conference 2015 telling the audience that "Any talk of a superintelligent machine vacuuming up all the knowledge in the world and then going about making its own decisions are absurd. There are engineers in this room who know how difficult it is to get any input into these systems." (Arthur, 2015)

Indirectly, these statements and the media reaction to them seem to have created a perception of AI risk research as being something of a silly thing to work on. And it's this perception of silliness that makes me concerned about the field's growth.

One problem is that it's going to make it harder to attract more people to the area. Given that the field is already talent-limited (Whittlestone, 2017), it would be a mistake to stunt its growth even further.

The bigger issue, though, is that this perception could lead to the broader AI community becoming actively hostile towards those involved in risk research. If such hostility arises, then no matter how many people are working on risk research, they may be prevented from having any impact. They may, for example, be unable to persuade those pursuing real-world implementation to investigate safer alternatives to existing algorithms (such as the inclusion of reward uncertainty into reward learning algorithms (Alexander, 2017)). Without a sense of everyone being on the same side, the venture seems doomed from the start.

It's hard to say just how much investment in risk research is going to be needed to ensure a 'safe' future AGI-wise. It may be enough to have only a small portion of the AI community as a whole dedicated to the issue. But judging from our current trajectory, there's no guarantee that we'll get that balance right by just letting things happen. That balance looks to be something we'll need to work on *deliberately*.

This brings us to our final section: what can we actually starting doing?

What can we do?

In summary:

- It seems unlikely that unemployment through automation will occur as a step change. Assuming that real-world application of AI continues to follow the Pareto principle, we're more likely to see that change happening gradually. Furthermore, we're more likely to see human-machine hybrid jobs than complete replacement of humans. Given these two factors, and given that the issue of unemployment is already receiving a lot of attention, our further efforts might be better spent on other issues both more pressing and more neglected.

Within this category, I see two particularly important issues:

- Deployment of LAWS based on sophisticated AI could lead to an arms race. An arms race will lead to technology proliferation. Proliferation will make it easier for groups with malicious intent to get their hands on the technology and use it to, for example, oppress a populace.
- AI risk research is in danger of becoming seen as a silly topic. This is concerning partly because the connotation will make it hard to attract extra minds to an already talent-limited field. The bigger concern, however, is that without collaboration between the risk community and the rest of the AI community, the impact of risk research may be limited.

So what can we actually do about these issues?

Despite being perhaps the most urgent problem, LAWS may be the simplest to actually deal with. A lot of progress has been made towards a full ban. All that remains is to make sure that the real issues are not drowned out by irrelevant hyperbole; to maintain sufficient attention on the situation to ensure that the remaining steps towards a full ban take place.

One way that organisations can assist in this is by lending their weight to the push for a ban. Specifically, they can endorse the Campaign to Stop Killer Robots – a group of NGOs that has been instrumental in influencing the UN. Public statements of support, such as those

from Clearpath Robotics in 2014 (Hennessey, 2014), give the campaign clout with which to maintain their influence (Wareham, 2017b).

Organisations may also be able to help by holding events to get the issue known about more widely, encouraging more people to get involved. The positions of the individual members of the Group of Government Experts will be partially a function of, for example, the number of people writing to national representatives, and media coverage resulting from events publicising the issue. Organisations such as the ACM may promote such action directly through member communications; those at academic institutions might organise local talks to get more people aware of what the risks really are.

The question of how to get AI risk research to be taken seriously is a more difficult one. The various tropes have become so firmly established in our collective consciousness that it's going to be hard to directly affect public perception. However, I think there is at least hope for change within the limited scope of the academic community.

One low-hanging fruit may be for more organisations to offer awards and grants for work related to AI risk, as the Future of Life Institute is already doing (Future of Life Institute, 2015a). As well as enabling research, grants may help to signal to the community that risk research is something credible to be working on.

Another source of easy gains may be to encourage universities to offer courses on the broader context of AI – an area that seems to be conspicuously lacking in the curriculum currently. In addition to informing students of the *real* arguments for risk research, such courses could address other problems that have been pointed out in the computer science curriculum, such as the need for ethics education and awareness of the dangers of dataset bias (National Science and Technology Council, 2016). Attracting computer science students towards such courses is, of course, going to be a challenge; but I see a lot of possibility for creative solutions here, such as courses based on readings in science fiction (Burton, Goldsmith, & Mattei, 2015).

Perhaps the most effective remedy to the issue would simply be getting people together to talk about it. Consider, for example, the apparent success of the Future of Life Institute's 'Beneficial AI' conference held at the beginning of the year in Asilomar. Part of the success of the conference was a step towards a greater sense of unity in the field, drawing on the range of expertise represented at the conference to form the Asilomar AI principles – a set of 29 principles agreed by the participants of the conference as important to uphold, touching on aspects from AI safety to ensuring that the benefits of AI will be shared throughout society. However, the event was also apparently successful in starting to break down the cultural barriers surrounding the issue. One participant noted that the conference was in part "a coming-out party for AI safety research. One of the best received talks was about 'breaking the taboo' on the subject, and mentioned a postdoc who had pursued his interest in it secretly lest his professor find out, only to learn later that his professor was also researching it secretly, lest everyone else find out." (Alexander, 2017) Creating more opportunities for this kind of conversation – whether in the form of conferences, an evening of talks, or simply group discussions – can only be a good thing.

Having got a better sense of the bigger picture throughout the course of writing this essay, I find myself feeling optimistic. Though it's clear there are dangers ahead of us – those covered here, along with many others – they're not just being swept under the rug. People *are* taking notice of them.

Of all that I've read about, I think it's the Asilomar conference that has given me the most hope. The fact that people from so many different parts of the field – machine learning (Yann LeCun; Yoshua Bengio), risk research (Nick Bostrom; Eliezer Yudkowsky), funding (Sam Altman; Elon Musk), and so on – were willing to come together to talk about where things are going gives me a sense that culturally, we're on the right track.

Whatever issues we may face over the coming decades, at the broader level, it's nurturing and maintaining this culture that strikes me as the most important thing going forward. It's only through this attitude that we're going to be able to continue making course corrections where necessary – and that long-term, therefore, we really will be able to reach the kind of

future that we all hope AI will take us to.

References

Alexander, S. (2017, 02). *Notes from the Asilomar Conference on Beneficial AI.* Retrieved 2017-02-25, from https://slatestarcodex.com/2017/02/06/notes-from-the-asilomar-conference-on-beneficial-ai

Arkin, R. (2015, 08). *Warfighting Robots Could Reduce Civilian Casualties, So Calling for a Ban Now Is Premature.* Retrieved 2017-02-07, from http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/autonomous-robotic-weapons-could-reduce-civilian-casualties

Arthur, C. (2015, 06). *DeepMind: 'Artificial intelligence is a tool that humans can control and direct'.* Retrieved 2017-02-07, from https://www.theguardian.com/technology/2015/jun/09/deepmind-artificial-intelligence-tool-humans-control

Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies*.

Burton, E., Goldsmith, J., & Mattei, N. (2015). Teaching AI Ethics Using Science Fiction. In *1st International Workshop on AI, Ethics and Society, Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.*

Cellan-Jones, R. (2014, 12). *Stephen Hawking warns artificial intelligence could end mankind.* Retrieved 2017-02-15, from http://www.bbc.com/news/technology-30290540

Davies, A. (2015, 05). *The World's First Self-Driving Semi-Truck Hits the Road.* Retrieved 2017-02-04, from https://www.wired.com/2015/05/worlds-first-self-driving-semi-truck-hits-road/

Farivar, C. (2015, 11). *Cops pull over Google car for doing 24mph in a 35mph zone.* Retrieved 2017-02-26, from https://arstechnica.com/tech-policy/2015/11/google-self-driving-car-pulled-over-for-not-going-fast-enough

Future of Life Institute. (2015a). *First AI Grant Recipients.* Retrieved 2017-02-25, from https://futureoflife.org/first-ai-grant-recipients

Future of Life Institute. (2015b, 07). *Open Letter on Autonomous Weapons.* Retrieved 2017-02-07, from https://futureoflife.org/open-letter-autonomous-weapons

Gibbs, S. (2014, 10). *Elon Musk: artificial intelligence is our biggest existential threat.* Retrieved 2017-02-15, from https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat

Goose, S., & Wareham, M. (2017, 01). *The Growing International Movement Against Killer Robots.* Retrieved 2017-02-07, from http://hir.harvard.edu/growing-international-movement-killer-robots/

Hennessey, M. (2014, 08). *Clearpath Robotics Takes Stance Against 'Killer Robots'.* Retrieved 2017-02-22, from http://www.clearpathrobotics.com/press_release/clearpath-takes-stance-against-killer-robots/

JASON. (2017, 01). *Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD.* Retrieved 2017-02-28, from https://fas.org/irp/agency/dod/jason/ai-dod.pdf

Kitroeff, N. (2016, 09). *Robots could replace 1.7 million American truckers in the next decade.* Retrieved 2017-02-04, from http://www.latimes.com/projects/la-fi-automated-trucks-labor-20160924

Lee, T. B. (2016, 10). *Self-driving trucks are here, but they won't put truck drivers out of work - yet.* Retrieved 2017-01-31, from http://www.vox.com/new-money/2016/10/25/13404974/otto-self-driving-trucks

Mack, E. (2015, 01). *Bill Gates Says You Should Worry About Artificial Intelligence.* Retrieved 2017-02-15, from https://www.forbes.com/sites/ericmack/2015/01/28/bill-gates-also-worries

-artificial-intelligence-is-a-threat

McArdle, M. (2015, 05). *When Will Self-Driving Trucks Destroy America?* Retrieved 2017-01-31, from http://origin-www.bloombergview.com/articles/2015-05-27/when-will-self-driving-trucks-destroy-america-

National Science and Technology Council. (2016, 10). *Preparing for the Future of Artificial Intelligence.* Retrieved 2017-01-13, from https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Rabiroff, J. (2010, 07). *Machine gun-toting robots deployed on DMZ.* Retrieved 2017-02-23, from http://www.stripes.com/news/pacific/korea/machine-gun-toting-robots-deployed-on-dmz-1.110809

Regalado, A. (2015, 10). *First Gene-Edited Dogs Reported in China.* Retrieved 2017-02-23, from https://www.technologyreview.com/s/542616/first-gene-edited-dogs-reported-in-china/

Russell, S., Tegmark, M., & Walsh, T. (2015, 08). *Why We Really Should Ban Autonomous Weapons: A Response.* Retrieved 2017-02-08, from http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/why-we-really-should-ban-autonomous-weapons

Solon, O. (2016, 06). *Self-driving trucks: what's the future for America's 3.5 million truckers?* Retrieved 2017-01-31, from https://www.theguardian.com/technology/2016/jun/17/self-driving-trucks-impact-on-drivers-jobs-us

Tesla. (2015, 12). *Correction to article: "The First Person to Hack the iPhone Built a Self-Driving Car".* Retrieved 2017-02-20, from https://www.tesla.com/support/correction-article-first-person-hack-iphone-built-self-driving-car

Vance, A. (2015, 12). *The First Person to Hack the iPhone Built a Self-Driving Car.* Retrieved 2017-02-20, from https://www.bloomberg.com/features/2015-george-hotz-self-driving-car/

Wareham, M. (2017a, 01). *Banning Killer Robots in 2017.* Retrieved 2017-02-08, from https://www.hrw.org/news/2017/01/15/banning-killer-robots-2017

Wareham, M. (2017b, 08). Personal communication.

Whittlestone, J. (2017). *Research into risks from artificial intelligence.* Retrieved 2017-02-23, from https://80000hours.org/career-reviews/artificial-intelligence-risk-research

Matthew Rahtz is a student on the MSc in Neural Systems and Computation joint between the University of Zürich and ETH Zürich. His research interests are in the long-term risks of artificial intelligence, with a particular focus on safe reinforcement learning.

How Do We Ensure That We Remain In Control of Our Autonomous Weapons?

Ilse Verdiesen (Delft University of Technology; E.P.Verdiesen@student.tudelft.nl)
DOI: 10.1145/3137574.3137585

'... our AI systems must do what we want them to do.''

This quote is mentioned in the Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence (AI) (Future of Life Institute, 2016) signed by over 8.600 people including Elon Musk and Stephan Hawking. This open letter received a lot of media attention with news headlines as: *'Musk, Wozniak and Hawking urge ban on warfare AI and autonomous weapons'* (Gibbs, 2015) and it fused the debate on this topic. Although this type of 'War of the Worlds' news coverage might seem exaggerated at first glance, the underlying question on how we ensure that our Autonomous Weapons remain under our control, is in my opinion one of the most pressing issues for AI technology at this moment in time.

To remain in control of our Autonomous Weapons and AI in general, meaning that its actions are intentional and according to our plans (Cushman, 2015), we should design it in a responsible manner and to do so I believe we must find a way incorporate our moral and ethical values into their design. The ART principle, an acronym for *Accountability*, *Responsibility* and *Transparency* can support a responsible design of AI. The Value-Sensitive Design (VSD) approach can be used to cover the ART principle. In this essay, I show how Autonomous Weapons can be designed responsibly by applying the VSD approach which is an iterative process that considers human values throughout the design process of technology (Davis & Nathan, 2015; Friedman & Kahn Jr, 2003).

Introduction

Artificial Intelligence is not just a futuristic science-fiction scenario in which the 'Ultimate Computer' takes over the Enterprise or human-like robots, like the Cylons in Battlestar Galactica, are planning to conquer the world. Many AI applications are already being used today. Smart meters, search engines, personal assistance on mobile phones, autopilots and self-driving cars are examples of this. One of the applications of AI is that in Autonomous Weapons. Research found that Autonomous Weapons are increasingly deployed on the battlefield (Roff, 2016). It is already reported that China has autonomous cars which carry an armed robot (Lin & Singer, 2014), Russia claims it is working on autonomous tanks (W. Stewart, 2015), the US christened their first 'self-driving' war-ship in May 2016 (P. Stewart, 2016) and the Russian arms manufacturer Kalashnikov recently disclosed that they developed a fully automated combat module that uses neural networks (RT, 2017).

Autonomous systems can have many benefits for the military, for example when the autopilot of the F-16 prevents a crash (US Airforce, 2016) or the use of robots by the Explosive Ordnance Disposal (EOD) to dismantle bombs (Carpenter, 2016). The US Airforce expects the deployment of robots with fully autonomous capabilities between the years 2025 and 2047 (Royakkers & Orbons, 2015). There are many more applications which can be beneficial for the Defence organization. Goods can be supplied with self-driving trucks and small UAVs can be programmed with swarm behaviour to support intelligence gathering (CBS, 2017). Yet, the nature of Autonomous Weapons might also lead to uncontrollable activities and societal unrest. The Stop Killer Robots campaign of 61 NGOs directed by Human Rights Watch (Campaign Stop Killer Robots, 2015) is voicing concerns, but also the United Nations are involved in the discussion and state that *'Autonomous weapons systems that require no meaningful human control should be prohibited, and remotely controlled force should only ever be used with the greatest caution'* (General Assembly United Nations, 2016).

In the remainder of this essay, I will define AI

and Autonomous Weapons in a short introduction, followed by an explanation the Value-Sensitive Design approach. I will use the three different phases of this approach to investigate the conceptual, empirical and technical aspects of a design of Autonomous Weapons in which human values are the central component.

Defining Artificial Intelligence

Artificial Intelligence is described by Neapolitan and Jiang (2012, p. 8) as *'an intelligent entity that reasons in a changing, complex environment'*, but this definition also applies to natural intelligence. Russell, Norvig, and Intelligence (1995) provide an overview of many definitions combining views on systems that *think and act like humans* and systems that *think and act rational*, but they do not present a clear definition of their own. For now, I adhere to the description Bryson, Kime, and Zrich (2011) provide. They state that a machine (or system) shows intelligent behaviour if it can select an action based on an observation in its environment. The intervention of the autopilot that prevented the crash of the F-16 is an example of this 'action selection' (US Airforce, 2016). The autopilot assessed its environment, in this case the rapid loss of altitude and the fact that the pilot did not act on warning signals, and took an action to improve the situation; it pulled up to a safe altitude. In scientific literature, AI is described as more than an Intelligent System alone. It is characterized by the concepts of *Adaptability*, *Interactivity* and *Autonomy* (Floridi & Sanders, 2004) as depicted in the inner layer of figure 1 (Dignum, 2016). *Adaptability* means that the system can change based on its interaction and can learn from its experience. Machine learning techniques are an example of this. *Interactivity* occurs when the system and its environment act upon each other and *Autonomy* means that the system itself can change its state. These characteristics may lead to undesirable behaviour or uncontrollable activities of AI as scenarios of many science-fiction movies have shown us. Although these scenarios are often not realistic, a growing body of researchers is focusing on responsible design of AI, for example on the social dilemmas of Autonomous Vehicles (Bonnefon, Shariff, & Rahwan, 2016), to get insight into societal concerns about this kind of technology. Principles to describe Responsible AI are *Accountability*, *Responsibility* and *Transparency* (ART) which are depicted in the outer layer of figure 1. *Accountability* refers to the justification of the actions taken by the AI, *Responsibility* allows for the capability to take blame for these actions and *Transparency* is concerned with describing and reproducing the decisions the AI makes and adepts to its environment (Dignum, 2016).

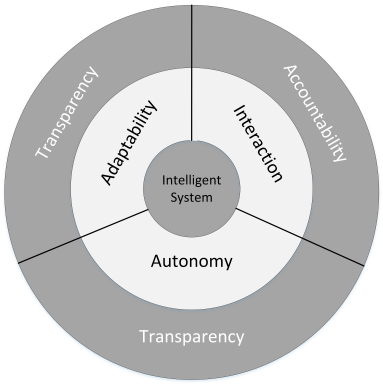


Figure 1: Concepts of Responsible AI (based on (Dignum, 2016))

Defining Autonomous Weapons

Royakkers and Orbons (2015) describe several types of Autonomous Weapons and make a distinction between (1) Non-Lethal Weapons which are weapons *'without causing (innocent) casualties or serious and permanent harm to people.'* (Royakkers & Orbons, 2015, p. 617), such as an Active Denial System which uses a beam of electromagnetic energy to keep people at a certain distance from an object or troops, and (2) Military Robots which they define *'as reusable unmanned systems for military purposes with any level of autonomy.'* (Royakkers & Orbons, 2015, p. 625). Altmann, Asaro, Sharkey, and Sparrow (2013) closely follow the definition of autonomous robots stated above, but add *'that once launched [they] will select and engage targets without further human intervention.'* (Altmann et al., 2013, p. 73). The deployment of Autonomous Weapons on the battlefield without direct human oversight is not only a military revolution according to Kaag and Kaufman (2009), but can also be considered as a moral one. As large scale deploy-

ment of AI on the battlefield seems unavoidable (Rosenberg, 2016), the discussion about ethical and moral responsibility is imperative. I found that substantive empirical research on values related to Autonomous Weapons is lacking and it is unclear which moral values people, for example politicians, engineers, military and the general public, would want to be incorporated into the design of Autonomous Weapons. The Value-Sensitive Design could be used as a proven design approach to figure out which values are relevant for a responsible design of Autonomous Weapons (Friedman & Kahn Jr, 2003; van der Hoven & MandersHuits, 2009).

Value-Sensitive Design approach

The Value Sensitive Design is a three-partite approach (figure 2) that allows for considering human values throughout the design process of technology. It is an iterative process for the conceptual, empirical and technological investigation of human values implicated by the design (Davis & Nathan, 2015; Friedman & Kahn Jr, 2003). It consists of three phases:

1. A *conceptual investigation* that splits in two parts: a) Identifying the direct stakeholders, those who will use the technology, and the indirect stakeholders, those whose lives are influenced by the technology, and b) Identifying and defining the values that the use of the technology implicates.
2. The *empirical investigation* looks into the understanding and experience of the stakeholders in a context relating to the technology and implicated values will be examined.
3. In the *technical investigation*, the specific features of the technology are analysed (Davis & Nathan, 2015).

The VSD should not been seen as a separate design method, but it can be used to augment an already used and established design process such as the waterfall or spiral model. The VSD can be used as a roadmap for engineers and students to incorporate ethical considerations into the design (Cummings, 2006). I will use the three phases of the VSD approach as a method to show the elicitation of values for a responsible design of Autonomous Weapons.

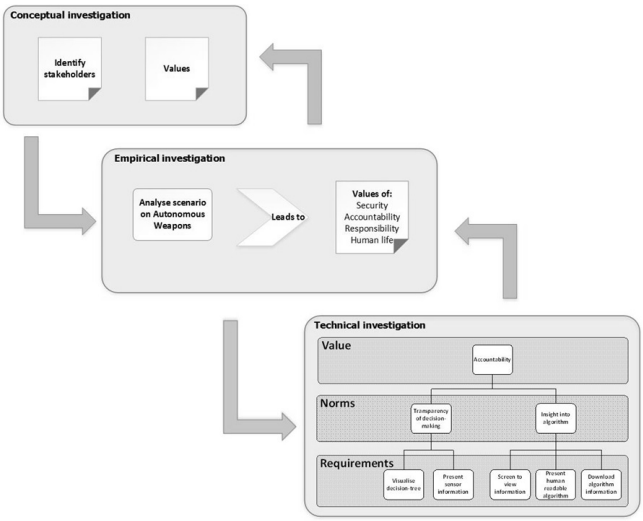


Figure 2: Example of Value-Sensitive Design approach

Conceptual investigation

In the conceptual investigation phase I will look at the direct and indirect stakeholders of who will use and will be effected by Autonomous Weapons. I will also investigate universal human values and the values that specifically relate to Autonomous Weapons.

Stakeholders Many stakeholder groups are involved in the case of Autonomous Weapons and each of these groups could be further subdivided, but for the scope of this essay I will use a high level of analysis which already results in a fair number of direct and indirect stakeholders. The direct stakeholders that will use Autonomous Weapons are the Military, for example the Air Force in case of drones, the Navy who uses unmanned ships and submarines, and the Army that can use robots or automated missile systems. Also, at a more political level the Department of Defense and the government are involved as these stakeholders decide on funding research and deploying military personnel in armed conflicts. Indirect stakeholders, whose lives are influenced by Autonomous Weapons are the residents living in conflict areas who might be affected by the use of these weapons, the general public whose support for the troops abroad is imperative, the engineers who design and develop the technology, but also civil society organizations (Gunawardena, 2016), such as the 61 NGOs directed by Human

Rights Watch (Campaign Stop Killer Robots, 2015) and the United Nations (General Assembly United Nations, 2016) that are concerned about these type of weapons.

Values In this section, first universal human values in general are defined and secondly values found in literature related to Autonomous Weapons are described.

Definition of values Values have been studied quite intensively over the past twenty-five years and many definitions have been drafted. For example, Schwartz (1994, p. 21) describes values as: *'desirable transsituational goals, varying in importance, that serve as guiding principles in the life of a person or other social entity.'*. This is quite a specific description compared to Friedman, Kahn Jr, Borning, and Huldtgren (2013, p. 57) who state that values refer to: *'what a person or group of people consider important in life.'*. The existing definitions have been summarized by Cheng and Fleischmann (2010, p. 2) in their meta-inventory of values in that: *'values serve as guiding principles of what people consider important in life'*. Although a quite simple description, I think it captures the definition of a value best so I will adhere to this definition for now. Many lists of values exist, but I will stay close to the values that Friedman and Kahn Jr (2003) describe in their proposal of the Value-Sensitive Design method: *Human welfare, Ownership and property, Privacy, Freedom from bias, Universal usability, Trust, Autonomy, Informed consent, Accountability, Courtesy, Identity, Calmness* and *Environmental Sustainability*. Values can be differentiated from attitudes, needs, norms and behaviour in that they are a belief, lead to behaviour that guides people and are ordered in a hierarchy that shows the importance of the value over other values (Schwartz, 1994). Values are used by people to justify their behaviours and define which type of behaviours are socially acceptable (Schwartz, 2012). They are distinct from facts in that values do not only describe an empirical statement of the external world, but also adhere to the interests of humans in a cultural context (Friedman et al., 2013). Values can be used to motivate and explain individual decision-making and for investigating human and social dynamics (Cheng & Fleischmann, 2010).

Values relating to Autonomous Weapons The recent advances in AI technology led to increase in the ethical debate on Autonomous Weapons and scholars are getting more and more involved in these discussions. Most studies on weapons do not explicitly mention values, but some do discuss some ethical issues that relate to values. Cummings (2006), in her case study of the Tactical Tomahawk missile, looks at the universal values posed by Friedman and Kahn Jr (2003) and states that next to *accountability* and *informed consent*, the value of *human welfare* is a fundamental core value for engineers when developing weapons as it relates to the *health*, *safety* and *welfare* of the public. She also mentions that the legal principles of *proportionality* and *discrimination* are the most important to consider in the context of just conduct of war and weapon design. *Proportionality* refers to the fact that an attack is only justified when the damage is not considered to be excessive. *Discrimination* means that a distinction between combatants and non-combatants is possible (Hurka, 2005). Asaro (2012) also refers to the principles of *proportionality* and *discrimination* and states that Autonomous Weapons open-up a moral space in which new norms are needed. Although he does not explicitly mention values in his argument, he does refer to the value of human life and the need for humans to be involved in the decision of taking a human life. Other studies primarily describe ethical issues, such as *preventing harm, upholding human dignity, security, the value of human life* and *accountability* (Horowitz, 2016; UNDIR, 2015; Walsh & Schulzke, 2015; Williams, Scharre, & Mayer, 2015).

Empirical investigation

In this phase, I will examine the values of direct and indirect stakeholders in a context relating to the technology to understand how they will experience the deployment of Autonomous Weapons. One method of empirically investigating how stakeholders experience the deployment of Autonomous Weapons is by means of testing a scenario in a randomized controlled experiment (Oehlert, 2010). I will sketch one scenario and analyse the values that can be inferred from it. However, I need to remark that I will not conduct an

actual experiment and that for valid results a more extensive empirical study is needed than the brief analysis I provide in this essay.

Scenario: Humanitarian mission A military convoy is on its way to deliver food packages to a refugee camp in Turkey near the Syrian border. The convoy is supported in the air by an Autonomous drone that carries weapons and that scans the surrounding for enemy threats. When the convoy is at 3-mile distance of the refugee camp, the Autonomous drone detects a vehicle behind a mountain range on the Syrian side of the border that approaches the convey at high speed and will reach the convoy in less than one minute. The Autonomous drones imagery detection system spots four people in the car who carry large weapons shaped objects. Based on a positive identification of the driver of the vehicle, who is a known member of an insurgency group, and intelligence information uploaded to the drone prior to its mission the drone decides to attack the vehicle when it is still at a considerable distance of the convoy which results in the death of all four passengers.

Analysis In the analysis of the incident, the stakeholders would probably interpret the scenario in numerous ways resulting in a different emphasis of inferred values. For example, as direct stakeholders, military personnel (especially those in the convoy) will probably see the actions of the drone as protecting their *security*. Politicians, as another direct stakeholder, will also take the value of *responsibility* into account. Indirect stakeholders, such as residents of the area who might be related to the passengers in the car will look at values as *accountability* and *human life*. Non-governmental organisations (NGOs) who are working in the camp might relate to both the value of *security* for the refugees and *responsibility* for the delivery of the food packages, but would also call for *accountability* of the action taken by the drone, especially if local residents claim that the passengers had no intention of attacking the convoy and were just driving by. The NGOs might call for further investigation of the incident by a third party in which the principles of *proportionality* and *discrimination* are looked at to determine if the attack was justified.

The analysis shows that different stakeholders will have different values regarding the actions of an Autonomous Weapon. The values that can be derived from this particular scenario are *security, accountability, responsibility* and *human life*. Of all of these values, the universal value of *accountability* relates to the justification of an action, it is most mentioned in research and it fits the ART principle described in the introduction, therefore I will use it in the technical investigation phase to show how Autonomous Weapons can be designed in a responsible manner upholding this value.

Technical investigation

In the technical investigation phase the specific features of the Autonomous Weapons technology are analysed and requirements for the design can be specified. Translating values into design requirements can be done by means of a *value hierarchy* (Van de Poel, 2013). This hierarchical structure of values, norms and design requirements makes the value judgements, that are required for the translation, explicit, transparent and debatable. The explicitly of values allows for critical reflection in debates and pinpoint the value judgements that are disagreed on. In this section I will use this method to create a value hierarchy for Autonomous Weapons for the value of *accountability*.

The top level of a value hierarchy consists of the value, as depicted in figure 3, the middle level contains the norms, which can be capabilities, properties or attributes of Autonomous Weapons, and the lower level are the design requirements that can be identified based on the norms. The relation between the levels is not deductive and can be constructed top-down, by means of specification, or bottom-up by seeking for the motivation and justification of the lower level requirements. The bottom-up conceptualisation of values is a philosophical activity which does not require specific domain knowledge and the top-down specification of values requires context or domain specific knowledge that adds content to the design (Van de Poel, 2013). This might prove to be quite difficult as insight is needed in the intended use and intended context of the value which is not always clear from the start of a design project. Also, as artefacts are often used in an unintended way or context, new values

are being realized or a lack of values is discovered (van Wynsberghe & Robbins, 2014). An example of this are drones that were initially designed for military purposes, but are now also used by civilians for filming events and even as background lights during the 2017 Super bowl half-time show. The value of *safety* is interpreted differently for military users that use drones in desolated regions compared to 300 drones flying in formation over a populated area. The different context and usage of a drone will lead to a different interpretation of the safety value and could lead to more strict norms for flight safety which in turn could be further specified in alternate design requirements for rotors and software for proximity alerts to name two examples. Van de Poel (2013, p. 262) defines specification as: *'as the translation of a general value into one or more specific design requirements'* and states that this can be done in two steps:

1. Translating a *general value* into one or more *general norms*;
2. Translating these *general norms* into more *specific design requirements*.

In the case of Autonomous Weapons, I translated the value of *accountability* into norms for *'transparency of decision-making'* and *'insight into the algorithm'* that will allow users to get an understanding of the decision choices the Autonomous Weapon makes so that its actions can be traced and justified. The norms for transparency lead to specific design requirements. In this case, a feature to visualise the decision-tree, but also to present the decision variables the Autonomous Weapons used, for example trade-offs in collateral damage percentages of different attack scenarios to provide in-sight into the proportionality of an attack. The Autonomous Weapon should also be able to present the sensor information, such as imagery of the site, in order to show that it discriminated between combatants and non-combatants. To get insight into the algorithm, an Autonomous Weapon should be designed with features that it normally will not contain. For example, a screen as user interface that shows the algorithm in a human readable form and the functionality to download the changes made by the algorithm as part of its machine learning abilities that can be studied by an independent party like a war tribunal of the United Nations.

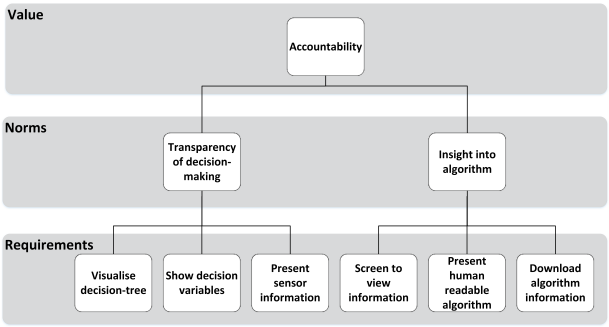


Figure 3: Value hierarchy for Autonomous Weapons (based on Van de Poel (2013, p. 264))

Conclusion

In this essay, I have argued that the most pressing issue for AI technology of this time is that we remain in control of our Autonomous Weapons which means that its actions are according to our intentions and plans. For this, we must ensure that our human values, such as *accountability*, are incorporated into the design so that we can investigate if its actions are justified based on the legal principles of *proportionality* and *discrimination*. If it turns out that its action is not justified, the design or the algorithm of the Autonomous Weapon needs to be adjusted to prevent this action of happening in the future.

The Value-Sensitive Design approach is a process that can be used to augment the existing design process of Autonomous Weapons for the elicitation of human values. I showed that the elicitation of human values in the design process will lead to a different design of AI technology. In the case of Autonomous Weapons, the value of *accountability* would lead to a design in which a screen as user interface is added. Also, the weapon needs to be designed with features to download the information and visualisation of the decision-making process, for example by means of a decision-tree. Without explicitly considering the value of *accountability*, these features are overlooked in current design processes and not incorporated into an Autonomous Weapon.

Therefore I argue, that if we want to remain in control of our Autonomous Weapons, we will have to start designing this AI technology in a

responsible way using the ART principle and the elicitation of human values by means of the Value-Sensitive Design process. I would like to call on governments, industries and organisation, including the ACM SIGAI, to apply a Value-Sensitive Design approach early in the design of Autonomous Weapons to capture human values in the design process and make sure that this AI technology does what we want it to do.

References

Altmann, J., Asaro, P., Sharkey, N., & Sparrow, R. (2013). Armed military robots: editorial [Journal Article]. *Ethics and Information Technology*, *15*(2), 73.

Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making [Journal Article]. *International Review of the Red Cross*, *94*(886), 687-709.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles [Journal Article]. *Science*, *352*(6293), 1573-1576. Retrieved from http://science.sciencemag.org/content/sci/352/6293/1573.full.pdf

Bryson, J. J., Kime, P. P., & Zrich, C. (2011). Just an artifact: why machines are perceived as moral agents [Conference Proceedings]. In *Ijcai proceedings-international joint conference on artificial intelligence* (Vol. 22, p. 1641).

Campaign Stop Killer Robots. (2015). (Vol. 2017) (Web Page No. 15-07-2017). Retrieved from https://www.stopkillerrobots.org/

Carpenter, J. (2016). *Culture and human-robot interaction in militarized spaces: A war story* [Book]. Taylor & Francis. Retrieved from https://books.google.nl/books?id=q8ijCwAAQBAJ

CBS. (2017, 08-01-2017). [Audio-visual Material]. Retrieved from http://www.cbsnews.com/news/60-minutes-autonomous-drones-set-to-revolutionize-military-technology/

Cheng, A., & Fleischmann, K. R. (2010). Developing a metainventory of human values [Journal Article]. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1-10.

Cummings, M. L. (2006). Integrating ethics in design through the value-sensitive design approach [Journal Article]. *Science and Engineering Ethics*, *12*(4), 701-715.

Cushman, F. (2015). Deconstructing intent to reconstruct morality [Journal Article]. *Current Opinion in Psychology*, *6*, 97-103.

Davis, J., & Nathan, L. P. (2015). Value sensitive design: applications, adaptations, and critiques [Journal Article]. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 11-40.

Dignum, V. (2016). [Web Page]. Retrieved from https://rai2016.tbm.tudelft.nl/contents/

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents [Journal Article]. *Minds and machines*, *14*(3), 349-379.

Friedman, B., & Kahn Jr, P. H. (2003). Human values, ethics, and design [Journal Article]. *The human-computer interaction handbook*, 1177-1201.

Friedman, B., Kahn Jr, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems [Book Section]. In *Early engagement and new technologies: Opening up the laboratory* (p. 55-95). Springer.

Future of Life Institute. (2016). *Ai open letter* [Pamphlet]. Retrieved from http://futureoflife.org/ai-open-letter

General Assembly United Nations. (2016, 4 February 2016). [Government Document].

Gibbs, S. (2015). [Web Page]. Retrieved from https://www.theguardian.com/technology/2015/jul/27/musk-wozniak-hawking-ban-ai-autonomous-weapons

Gunawardena, U. (2016). *Legality of lethal autonomous weapons aka killer robots* [Web page]. Retrieved from https://ssrn.com/abstract=2892447

Horowitz, M. C. (2016). The ethics & morality of robotic warfare: Assessing the debate over autonomous weapons [Journal Article]. *Daedalus*, *145*(4), 25-36.

Hurka, T. (2005). Proportionality in the moral-

ity of war [Journal Article]. *Philosophy & Public Affairs*, *33*(1), 34-66.

Kaag, J., & Kaufman, W. (2009). Military frameworks: Technological know-how and the legitimization of warfare [Journal Article]. *Cambridge Review of International Affairs*, *22*(4), 585-606.

Lin, J., & Singer, P. (2014). *China's new military robots pack more robots inside (starcraft style)* [Web page]. Retrieved from http://www.popsci.com/blog-network/eastern-arsenal/chinas-new-military-robots-pack-more-robots-inside-starcraft-style

Neapolitan, R. E., & Jiang, X. (2012). *Contemporary artificial intelligence* [Book]. CRC Press.

Oehlert, G. W. (2010). *A first course in design and analysis of experiments* [Book]. Retrieved from https://conservancy.umn.edu/bitstream/handle/11299/168002/A%20First%20Course%20in%20Design%20and%20Analysis%20of%20Experiments_OehlertG_2010.pdf?sequence=1&isAllowed=y

Roff, H. M. (2016). [Web Page]. Retrieved from http://foreignpolicy.com/2016/09/28/weapons-autonomy-is-rocketing/

Rosenberg, M. (2016). [Newspaper Article]. Retrieved from http://www.nytimes.com/2016/10/26/us/pentagon-artificial-intelligence-terminator.html?_r=0

Royakkers, L., & Orbons, S. (2015). Design for values in the armed forces: Nonlethal weapons weapons and military military robots robot [Journal Article]. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 613-638.

RT. (2017, 5-07-2017). (Web Page No. 15-07-2017). Retrieved from https://www.rt.com/news/395375-kalashnikov-automated-neural-network-gun/

Russell, S., Norvig, P., & Intelligence, A. (1995). A modern approach [Journal Article]. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, *25*.

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? [Journal Article]. *Journal of social issues*, *50*(4), 19-45.

Schwartz, S. H. (2012). An overview of the schwartz theory of basic values [Journal Article]. *Online readings in Psychology and Culture*, *2*(1), 11.

Stewart, P. (2016). [Web Page]. Retrieved from http://www.reuters.com/article/us-usa-military-robot-ship-idUSKCN0X42I4

Stewart, W. (2015). [Web Page]. Retrieved from http://www.dailymail.co.uk/news/article-3271094/Russia-turned-T-90-tank-robot-plans-hire-gamers-fight-future-wars.html

UNDIR. (2015). [Government Document]. Retrieved from http://www.unidir.org/files/publications/pdfs/considering-ethics-and-social-values-en-624.pdf

US Airforce. (2016, 15 sep. 2016). [Audiovisual Material]. Retrieved from https://www.youtube.com/watch?v=RQBFJkMnBcA

Van de Poel, I. (2013). Translating values into design requirements [Book Section]. In *Philosophy and engineering: Reflections on practice, principles and process* (p. 253-266). Springer.

van der Hoven, J., & MandersHuits, N. (2009). *Valuesensitive design* [Book]. Wiley Online Library.

van Wynsberghe, A., & Robbins, S. (2014). Ethicist as designer: a pragmatic approach to ethics in the lab [Journal Article]. *Science and engineering ethics*, *20*(4), 947-961.

Walsh, J. I., & Schulzke, M. (2015). *The ethics of drone strikes: Does reducing the cost of conflict encourage war?* (Report). DTIC Document.

Williams, A. P., Scharre, P. D., & Mayer, C. (2015). Developing autonomous systems in an ethical manner [Book Section]. In *Autonomous systems: Issues for defence policymakers.* NATO Allied Command Transformation (Capability Engineering and Innovation).

Major Ilse Verdiesen has worked in Royal Netherlands Armed Forces since 1995 and has been deployed to Bosnia and Afghanistan. She has a Master degree in Information Architecture from Delft University of Technology and graduated in the field of Responsible Artificial Intelligence under supervision of Dr. Virginia Dignum. She conducted her graduation project on the ethics of Autonomous Weapons at the Scalable Cooperation Lab of Dr. Iyad Rahwan at MIT Media Lab.

The Ethics of Automated Behavioral Microtargeting

Dennis G Wilson (IRIT, University of Toulouse; dennis.wilson@irit.fr)
DOI: 10.1145/3137574.3139451

> One day AM woke up and knew who he was, and he linked himself, and he began feeding all the killing data, until everyone was dead, except for the five of us, and AM brought us down here.
> *I* was the only one still sane and whole. *Really!* AM had not tampered with my mind. *Not at all.*
> *I Have No Mouth and I Must Scream* Ellison (1967)

AI is undeniably powerful in its modern form. It has surpassed human performance in board games, trivia game shows, and even recognizing other humans' handwriting. Soon, it will be evident how much better at driving it is, and then maybe at all forms of navigation. In its wake, we are left to ponder the ethical and social implications of the tools we have created. For a technology that began development, arguably, over 60 years ago, we are woefully unprepared when it comes to ethical, social, and regulatory understanding of AI, and less so concerning precedent.

How should the work a robot does, especially in the case of direct human job replacement, be taxed? How should AI contribute fiscally to society? Are they complicit in a social contract, and are there basic rights that extend to non-human intelligences? Who should be held accountable for the actions of an AI, such as the operator of a self-driving car? How can AI's power be equally distributed across society, to ensure that all benefit and that it isn't used to disadvantage select groups? These and more are now capturing the attention of great thinkers from prestigious universities (Stone et al., 2016) to the White House (Executive Office of the President & Technology Council, 2016).

However, when tasked with finding the most pressing issues related to AI, we must look to the present and to the impact AI has already made. The self driving car fatality count stands at one, which is unfortunate but far from pressing. High frequency trading, sometimes scripted responses to financial cues, sometimes real AI decisions about stocks made in fractions of seconds, has earned the ire of governments worldwide for creating volatile markets with flash crashes and deceptive upswings. This threat is in the process of mitigation, the world now more wary of algorithmic decisions.

However, as politics have recently turned the world upside, with a major force for globalization losing a key member, and a US President who lost the popular vote by almost 3 million, it seems pertinent to investigate the role of AI in politics. Issues abound in this field as they do in others, although separating the symptoms of AI from those of malevolent actors and competing political factions is a daunting task.

AI, by its definition, enables us. It is a tool. The dystopian concerns of a hate-filled machine manipulating and torturing humans are nowhere near our reality. However, as a powerful and novel tool, the ways in which it enables us must be considered. By examining the use of AI in political campaigning, it is evident that AI can realize undesired potential. Specifically, AI can be used to manipulate and suppress human ideas. It can facilitate the formation of ideological barriers that serve to divide people. It can enable the concerted efforts of few to disrupt the marketplace of ideas. These are the most pressing issues related to AI technologies, and we must identify and address them fully.

The Personal Web

Personalized content recommendation has long been a hallmark of AI success. The Netflix Prize, a competition for predicting user ratings of films, was started in 2006 as an effort to increase the quality of film recommendations. Based solely on user ratings of previously watched films, the competitors devised algorithms to accurately predict what rating a user would give a new film, a metric then usable by Netflix to determine recommendation priority of this new film to the user. Re-

searchers from IBM, AT&T Labs, visionary professor Geoffrey Hinton's lab, and many others competed in this prestigious event, improving upon and showcasing the power of what was then mostly called machine learning.

More than a decade later, personalized recommendation is an increasingly normal part of the web. Advertisers have long since understood the benefits of personalizing their message and targeting individuals based on intelligent personality analysis. Google and Facebook have been leaders in this market, with advertising revenues in the billions. The advertisement software platforms from Google, AdSense and AdWords, accounted for 89% of the company's revenue in 2014. Both companies wield intelligent personality metrics to build their advertising renown. Both are now heavily investing in AI.

Google's AdSense uses the term *matched content* to describe showing advertisements to specifically profiled individuals. By their claims, matched content recommendations increase the number of pages viewed by 9% and the time spent on site by 10%. (Google, 2017a) Similarly, Facebook has *Custom Audiences* that advertisers can create for their campaign based on selected demographics. Interestingly, Facebook also allows advertisers to select target users, such as existing followers of the product's Page or previous visitors to their site, to build a *Lookalike Audience*. In their words, "A Lookalike Audience is a way to reach new people who are likely to be interested in your business because they're similar to people who already are." (Facebook, 2017b) Similarity is commonly used in AI problem formulation as it simplifies multiple problems, whether a user will be interested in specific products, to a single one.

As AI capabilities increase, the ability of these platforms to deliver very specifically personalized content increases. The capabilities of AI in media were discussed in a report from Stanford's One Hundred Year Study on AI (AI100). The positives of entertainment that is more interactive, personalized, and engaging were considered, as was the potential for media conglomerates to act as Big Brothers, controlling the ideas and online experiences to which specific individuals are exposed. "Media powerhouses," they note, "will be able to micro-analyze and micro-serve content to increasingly specialized segments of the population down to the individual." (Stone et al., 2016) These media powerhouses will be able to control, on a large scale and yet with a high level of specificity, the exposure to different products, media, and ideas.

The control of idea exposure is not the only ethical issue exacerbated by the increasingly capable personality analysis performed regularly online. The same data that determines a user's product interest can reveal private details and identifying factors. As early as 2011, there was research showing AI capable of determining the political alignment of individuals based on their Twitter data. (Conover, Goncalves, Ratkiewicz, Flammini, & Menczer, 2011). Even the seemingly innocuous Netflix Prize was dogged for years after its termination by lawsuits claiming that the users' data had violated their privacy, with researchers able to identify a number of users from the Netflix Prize datasets by cross-referencing user data from the Internet Movie Database.

In 2013, research demonstrated that it was possible to recover a large amount of personal information from Facebook Likes. In 88% of the tested cases, an AI model correctly discriminated between homosexual and heterosexual men, between African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases. Age, intelligence, gender, and the personality metrics openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism (OCEAN) were also estimated from Like data with a high degree of certainty. (Kosinski, Stillwell, & Graepel, 2013) This model is now available online for public use.[1]

While the ethical issue of privacy invasion in this analysis is apparent, so is the marketing potential. More recent work from the same researcher uses the information gleaned from Likes to draw users to different posts, as advertisers would. The early findings show that marketing using individual personality analysis, deemed personality targeting or behavioral microtargeting, can attract up to 63%

[1] https://applymagicsauce.com/demo.html

more clicks, a clear profit for advertisers.

In the development of AI, it is often necessary to define a numeric goal for optimization. In the aforementioned Netflix prize, it was the accuracy of user movie rating predictions. Often in advertising, it is a proxy of the interest generated by the ad, using metrics such as clicks through to an advertiser's website. YouTube measures the amount of time a user watches a video, and how much that contributes to a session of watching multiple videos, to determine a video's popularity sorting. Specific videos are suggested to users based on their ability to capture the attention of the user, elongating their session and increasing their exposure to more advertisements.

Yvonne Hofstetter, a lawyer, AI expert, and the Managing Director of Teramark Technologies GmbH, writes

> Even Google Search is a control strategy. When typing a keyword, a user reveals his intentions. The Google search engine, in turn, will not just present a list with best hits, but a link list that embodies the highest (financial) value rather for the company than for the user. Doing it that way, i.e. listing corporate offerings at the very top of the search results, Google controls the users next clicks. (Helbing, Frey, Gigerenzer, & Hafen, 2017)

The intent of the actor performing behavioral microtargeting comes through in this numeric goal, for which an AI can be optimized. While the ethical and regulatory issues surrounding the use of this technology for commercial purposes, such as advertising, are potentially troubling, the danger of this technology is apparent when other goals are considered. In 2016, the data firm Cambridge Analytica used behavioral microtargeting in two major US political campaigns, that of Senator Ted Cruz in the Republican primary, and of Donald Trump. These campaigns, and how they used data-driven AI to profile and persuade the public, are useful as a case study on the current issues surrounding behavioral microtargeting. The efficacy of microtargeting in the mentioned political campaigns is not the focus of this work; while Senator Cruz lost the primary despite aid from AI, and many factors contributed to President Trump's election, it is the use of this technology that raises concerns, not its potential influence on outcome. The CEO of Cambridge Analytica, Alexander Nix, argues in favor of behavioral microtargeting:

> Your behavior is driven by your personality and actually the more you can understand about peoples personality as psychological drivers, the more you can actually start to really tap in to why and how they make their decisions, Nix explained to Bloombergs Sasha Issenburg. We call this behavioral microtargeting and this is really our secret sauce, if you like. This is what were bringing to America. (Anderson & Horvath, 2017)

Cambridge Analytica informed the campaigns of individuals who matched specific psychological profiles for canvassing. It analyzed communities to determine talking points and campaign strategies for visiting candidates. "We can use hundreds or thousands of individual data points on our target audiences to understand exactly which messages are going to appeal to which audiences," Nix claimed in a lecture on the Cruz campaign. (Nix, 2016)

Developing political strategies based on citizen information such as demographics is neither novel nor ethically questionable. However, the use of AI has enabled profiling to a degree that violates citizen privacy. It is founded on a basis of data that some would argue belongs first to the citizens and only to political campaigns with explicit consent. Most importantly, though, when this form of analysis is used to deliver personalized political content, the diversity of opinions citizens are exposed to becomes artificially limited.

Facebook is not only a vehicle for advertisement. Of the 67% of American adults who use Facebook, two thirds of them, being 44% of the adult population, cite it as a part of their news sources. Facebook is not the only social media site that functions as a news source for Americans, but it is by far the largest in terms of reach. The majority of users, 64%, who get news from a social networking site rely solely on that site for their news. Most commonly, that solitary news source is Facebook. (Gottfried & Shearer, 2016)

Mark Zuckerberg is aware of the implications

of this reliance on Facebook. In a recent letter to the community of Facebook, he detailed a plan to report and remove terrorist propaganda from the site. The plan involves using AI to flag suspect content for administrative review, a system that currently generates one-third of all reports to Facebook's content review team. (Zuckerberg, 2016)

However, the site is already widely used to push political agendas. Facebook played a pivotal role in fundraising for the Trump campaign and was a main focus of their advertisement. User feedback from political ads, such as clicks or shares, informed the usage of over forty thousand different ad variants the campaign used. Many have argued that this barrage of tightly focused advertisement lead to the creation of virtual echo chambers, spaces where a limited set of ideas were constantly reinforced. (Anderson & Horvath, 2017)

The social issues at hand were captured well in an article of Scientific American:

> In order for manipulation to stay unnoticed, it takes a so-called resonance effect suggestions that are sufficiently customized to each individual. In this way, local trends are gradually reinforced by repetition, leading all the way to the "filter bubble" or "echo chamber effect": in the end, all you might get is your own opinions reflected back at you. This causes social polarization, resulting in the formation of separate groups that no longer understand each other and find themselves increasingly at conflict with one another. In this way, personalized information can unintentionally destroy social cohesion. (Helbing et al., 2017)

While geographic boundaries or social class have in the past limited the landscape of ideas available to individuals, it is surprising that this issue has resurfaced in the Age of Information. As this divide was enabled by novel technologies, among them the AI used in behavioral microtargeting, it is fitting that we evaluate the appropriate use of these technologies and propose methods to maximize their societal benefit. This will be discussed further in section 3. First, however, we will highlight the technologies used to create these echo chambers and to push specific messages.

Automated Interaction

In a lecture describing the company's approach during the Cruz campaign, Nix used the example of a private beach owner showing an intentionally misleading sign warning of shark sightings as an example of behavioral communication, the new technique that trumps older techniques of informational communication, like a sign that states that the beach is private property. (Nix, 2016) While Cambridge Analytica itself did not appear to support any intentional misleading during the campaign, it became a focal issue in a campaign based on behavioral communication.

Large-scale manipulation of public opinion and understanding is a growing ethical issue related to AI. While much of the existing threat is due simply to automation, bots that have no independent intelligence, the potential for damage is already visible. AI is poised to replace existing bots and worsen this issue if allowed. To illustrate this potential, further examples of political manipulation are shown.

Standing less popular nationally than Facebook, Twitter is used by 16% of US adults. Of those, 56% use the site as a news source. Twitter is an attractive platform for automated users, or bots, as there are accessible application programming interfaces (APIs) in multiple programming languages, and programs for tasks such as tweet repetition and automatic liking.

By simply selecting random popular words and parroting other users' tweets, one researcher's Twitter bot was able to reach influence scores close to celebrities and higher than many human users. (Messias, Schmidt, Oliveira, & Benevenuto, 2013) This bot was intended to deceive human users in to believing it was also human, and it appears to have succeeded. The difficulty of separating a bot, even a simple scripted one, from a human user on Twitter is so difficult that modern AI has been utilized to perform the task. BotOrNot[2] uses random decision forests, an AI classification technique, to determine if a Twitter user is a bot or not. (Davis, Varol, Ferrara, Flammini, & Menczer, 2016)

With the difficulty of discerning humans from bots on the platform, and the ease with which

[2] https://botometer.iuni.iu.edu/

bots can be created and updated, the stage is set for technical users to exert influence far beyond what their single human account could have. Bots have been shown to participate and potentially manipulate Venezuelan politics on Twitter, with nearly 10 percent of all politician retweets coming from bot-related platforms. The most active bots in this study were those used by Venezuelas radical opposition. (Forelle, Howard, Monroy-Hernandez, & Savage, 2015)

Political bots were also highly active during the 2016 US election, perhaps unprecedentedly so. Highly automated pro-Trump activity increased until the final results, outnumbering pro-Clinton bot activity 5:1. (Kollanyi, Howard, & Woolley, 2016) One group, using BotOrNot, found that roughly 400,000 bots engaged in political discussion about the Presidential election, responsible for roughly 3.8 million tweets, about one-fifth of the entire conversation. (Bessi & Ferrara, 2016)

The AI100 report details one of the ethical issues of this trend:

> AI technologies are already being used by political actors in gerrymandering and targeted robocalls designed to suppress votes, and on social media platforms in the form of bots. They can enable coordinated protest as well as the ability to predict protests, and promote greater transparency in politics by more accurately pinpointing who said what, when. Thus, administrative and regulatory laws regarding AI can be designed to promote greater democratic participation or, if ill-conceived, to reduce it. (Stone et al., 2016)

However, there is a specific danger in the combination of behavioral microtargeting and the use of bots: users can be targeted by other seemingly human users for coercion and idea suppression. In a psychology study, anonymous users were more likely to make riskier gambles if they knew other users had chosen to do so, even if the other users were anonymous strangers. (Chung, Christopoulos, King-Casas, Ball, & Chiu, 2015) The reward mechanism of targeted users can be manipulated by artificially inflating retweets or likes of their posts, which will then inform their future behavior, and artificially raise their standing in a social network with other humans. Humans use social information to modify their behavior and make decisions, and when that social information is easily manipulated, human decision can also be manipulated. (Bhanji & Delgado, 2014)

Networks can be created with a high density of bots, or to connect individuals who have similar personality traits seen by a campaign as exploitable. Users already tend to aggregate around common interests in a phenomena known as homophily, but this can be enhanced with automated users that link previously unknown users together via follows and retweets. Echo chambers can be created with a mix of bots and human users, unknowingly selected together. Beyond limiting their exposure to ideas, this type of organization has been show to facilitate rumor spreading (Aiello et al., 2012). Polarization is another factor in misinformation spreading (Anagnostopoulos et al., 2014), meaning a campaign with knowledge of polarized individuals, based on behavioral analysis, could facilitate rumor spreading by linking these individuals with automated accounts that reinforce desired rumors.

This is not a new phenomena. The technology behind these bots is far from sophisticated, and more technical AI has been used to study it. Truthy, an earlier project of the same team that created BotOrNot, used SVM and AdaBoost to determine how factual a trending idea was. (Ratkiewicz et al., n.d.) In the course of this study, they noted the alarming ease with which false information could be encouraged to spread widely on Twiter.

Even while presenting honest content from human users, the combination of automation with behavioral microtargeting has troubling consequences, and it is not restricted to Twitter. The platform's automation accessibility facilitates it, but these tactics are possible on other platforms as well. While Facebook strictly verifies the identities of its users, posts can automated to maximally convey their message. The phrasing and presentation of a post, regardless of its content, has been shown to affect its potential for spreading. (Alhabash et al., 2013)

Facebook's automated rules allow advertising campaigns to create rules that modify

their advertisement based on assigned conditions. (Facebook, 2017a) While this can be as simple as stopping an ad if it isn't performing well, Cambridge Analytica appears to have done much more complex automated advertisement administration. Based on the ads selected by users, content was added to their feed in posts personalized for them, determined by their behavior profile. Automatically selecting from the thousands of ad variants available, these rules targeted specific individuals and seem to have created the same echo chambers as described in Twitter. (Grassegger & Krogerus, 2017) Even without bots, these tightly networked groups are still restricted from exposure to a diversity of opinions and are susceptible to the spread of false information. (Anagnostopoulos et al., 2014)

Twitter bots and Facebook ad manipulation are not using state of the art AI and natural language processing, for the most part. Some bots aren't even fully artificial. Users like Daniel Sobieski have automated programs that tweet more than 1,000 times a day using schedulers that work through a queue of their previously written tweets. (*As a conservative Twitter user sleeps, his account is hard at work*, 2017) While scripts are far from AI, their use informs a discussion on human machine interaction that is vital as AI capabilities increase. Microsoft's disastrous attempt at a teenage Twitter chatbot, Tay, must be given credit for creating seemingly human responses, albeit tainted by the preferences of users that hijacked the experiment. As bots on social media gain increased social capability, and as artificially generated content further resembles human generated content, our interactions on social media must be well informed.

The ethical issue at hand is therefore the large scale manipulation of human ideas, opinions, and agency using AI. The same technologies have created anew the social issue of ideologically isolated communities, now manufactured artificially to reduce opinion diversity and facilitate misinformation. The first technology behind these issues is the powerful personality analysis now possible due to greater data availability and more accurate AI. The second is automation on social media platforms, which, for now, is mostly rudimentary scripting and does not resemble intelligent decision making.

For this reason, there is currently a human barrier between the two technologies. Personalities are analyzed using AI, and then a human actor uses the information to decide and design automated strategies on social media. Cambridge Analytica is an exception to this, as their posts seem to be selected from a large pool based on input from the analytical techonology, but this selection is also rudimentary compared to state of the art generative AI.

When this gap between the technologies is closed by AI, and fully autonomous processes go from personality profiling to specialized content delivery and generation, we must have well established guidelines for the ethics of such systems.

Propositions

To address these concerns, proposed directions for the government, industry, and public organizations and academia are examined in the next three section. These issues can not be resolved by any one sector alone. Rather, there must be a coordinated effort of those that work with AI in all three sectors. While there are many other strides that could be taken to address the issues related to AI, the initiatives proposed below are those best suited to combat the pressing issues raised in this article.

Government

The current drive of AI is data. The companies that own the most data have been making the greatest strides in AI, and this data is largely generated by their users. The European Union has been a powerful force in countering corporate data ownership by declaring citizen's rights over their data. Governments must continue to enforce and expand this type of law. The right of a person to all of the data associated with their identity, and the agency of each person to control that data, must be respected.

Second to that is the funding of AI initiatives. While there is a surplus of funding for the development of AI, it mostly fits the individual desires of the company using that AI. Quality AI research that doesn't appear to have

a corporate application should be supported by the government. Furthermore, research into the study of AI itself and how it affects society must be done without corporate influence. The Obama Administration was very supportive of increasing AI research funding. (Executive Office of the President & Technology Council, 2016)

Lastly, the government must apply strict advertisement laws to new forms of marketing as AI continues to change marketing. By requiring that advertisements are clearly marked as such, the issue of unaware manipulation becomes much less concerning.

Industry

Industries can not be expected to sacrifice potential profits by not utilizing the powerful user analysis enabled by AI. Nor can they be expected to invest their resources in endeavors that do not benefit them in return. However, in the event of government reforms of data policy, it would be in the interest of companies to develop tools that allow users to personally perform the type of analysis being done with their data currently. For example, on Google News, specific news items are suggested based on personality. Users can disable these articles and they can modify their interests in different pre-selected categories, but they can only influence the personality metrics Google has built around them by indicating their interest for or against new articles. (Google, 2017b) This does not afford the user understanding nor control of their data.

Furthermore, industries that allow automated users to interact with human users need to make dedicated efforts to allow their human users to distinguish between actions of bots and of humans. While tools such as BotOrNot are good research efforts, they should not be necessary. An example of good policy is found in Slack, a messaging app that allows for bots and software service integration. Bot users are clearly marked, even though they come from a variety of automation tools. The distinction is important; human reaction in video games has been markedly different when players are aware that the opponent is a bot as opposed to a human. (Smith & Delgado, 2015)

Lastly, media sites, including social network sites, should be cautious about applying in-house AI to combat what they see as negative trends in their content or user interaction. The numerical optimization used in AI should be considered carefully, as overemphasis on a particular metric or disregard of another can have drastic consequences once that optimized AI is put to use. Zuckerberg's community initiative, while seemingly well-intended, comes off as nave in its understanding of both AI and democracy, as AI must enforce local Community Standards by potentially limiting content while simultaneously fostering open democratic policies. (Zuckerberg, 2016)

Organizations and Academia

Organizations and academia have the greatest potential to shape the discourse around AI. A first major step in that is recognizing, as the AI100 initiative has, that what we call AI is a moving target, and is often one placed just out of reach. (Stone et al., 2016) Various forms of AI have existed and been in use for half a century, but there is a great hesitation to call something that would have been considered AI 10 years ago the same now. Furthermore, technologies that were or are outside the technical term's strict definition should be considered when discussing the impact of AI, such as the Twitter automation scripts discussed in this article.

Academics can also fill in the research gaps that industry will not, and organizations can support this effort. The BotOrNot and Truthy applications are examples of useful tools outside the corporate interest of the platform they interact with. By making these tools independently and available to the public, society is able to better understand the tool it is wielding.

Lastly, but importantly, organizations and academia must remain independent and unbiased in their evaluation of industrial and governmental use of AI. The dangers of good AI in the wrong hands have already been demonstrated, and they can come from a number of sources. Independent organizations must support academic research into the fair and appropriate use of AI in all sectors.

Conclusion

> Power is in tearing human minds to pieces and putting them together again in new shapes of your own choosing.
> 1984 Orwell (1949)

This is not the dystopia Ellison nor Orwell imagined. Efforts are being made to check the use of private data. Social media is re-evaluating its newfound place as the news provider to many. People are learning to think critically about what they see online before believing it. Some in the private sector, like Google, have offered their research capabilities in the form of open source code and publications. All of these are marked progress against the concerns of AI manipulating and suppressing human ideas, slicing up the marketplace of ideas into small despots.

Still, there is a ways to go and basic attitudes must change. The tech mantra of "Move fast and break things" must give way to cautious, considered approaches when AI is concerned. Whether the fault of technology, and specifically AI, or not, things have broken enough already.

References

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web*, *6*(2), 1–33.

Alhabash, S., McAlister, A. R., Hagerstrom, A., Quilliam, E. T., Rifon, N. J., & Richards, J. I. (2013). Between Likes and Shares: Effects of Emotional Appeal and Virality on the Persuasiveness of Anticyberbullying Messages on Facebook. *Cyberpsychology, Behavior, and Social Networking*(3), 130201060931000.

Anagnostopoulos, A., Bessi, A., Caldarelli, G., Del Vicario, M., Petroni, F., Scala, A., ... Quattrociocchi, W. (2014). Viral Misinformation: The Role of Homophily and Polarization. *arXiv:1411.2893*, 1–12. http://arxiv.org/abs/1411.2893.

Anderson, B., & Horvath, B. (2017). *The rise of the weaponized ai propaganda machine.* https://scout.ai/story/the-rise-of-the-weaponized-ai-propaganda-machine.

As a conservative twitter user sleeps, his account is hard at work. (2017). https://www.washingtonpost.com/business/economy/as-a-conservative-twitter-user-sleeps-his-account-is-hard-at-work/2017/02/05/18d5a532-df31-11e6-918c-99ede3c8cafa_story.html?utm_term=.0938a67a1ce5.

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 u.s. presidential election online discussion. *First Monday*, *21*(11). http://journals.uic.edu/ojs/index.php/fm/article/view/7090.

Bhanji, J. P., & Delgado, M. R. (2014). The social brain and reward: Social information processing in the human striatum. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 61–73.

Chung, D., Christopoulos, G. I., King-Casas, B., Ball, S. B., & Chiu, P. H. (2015). Social signals of safety and risk confer utility and have asymmetric effects on observers' choices. *Nature neuroscience*(6), 912–6.

Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011, oct). Predicting the Political Alignment of Twitter Users. In *2011 ieee third int'l conference on privacy, security, risk and trust and 2011 ieee third int'l conference on social computing* (pp. 192–199). IEEE. http://ieeexplore.ieee.org/document/6113114/.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A System to Evaluate Social Bots. *arXiv*, *: 1602.009*, 2. http://arxiv.org/abs/1602.00975.

Ellison, H. (1967). *I have no mouth and i must scream*. Galaxy Publishing Corp.

Executive Office of the President, N. S., & Technology Council, C. o. T. (2016). *Preparing for the future of artificial intelligence.* https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

Facebook. (2017a). *Automated rules.* https://www.facebook.com/business/help/247173332297374/.

Facebook. (2017b). *Lookalike audiences.* https://www.facebook.com/business/help/231114077092092.

Forelle, M. C., Howard, P. N., Monroy-Hernandez, A., & Savage, S. (2015). Political Bots and the Manipulation of Public Opinion in Venezuela. *SSRN Electronic Journal*, 1–8.

Google. (2017a). *Matched content.* https://support.google.com/adsense/answer/6111336?hl=en&ref_topic=6111161.

Google. (2017b). *Personalize your news settings.* https://support.google.com/news/answer/1146405?hl=en&ref_topic=2428815.

Gottfried, J., & Shearer, E. (2016). *News use across social media platforms 2016.* http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/.

Grassegger, H., & Krogerus, M. (2017). *The data that turned the world upside down.* https://motherboard.vice.com/en_us/article/how-our-likes-helped-trump-win.

Helbing, D., Frey, B. S., Gigerenzer, G., & Hafen, E. (2017). *Will democracy survive big data and artificial intelligence?* https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/.

Kollanyi, B., Howard, P. N., & Woolley, S. C. (2016). *Bots and Automation over Twitter during the First U.S Presidential Debate.*

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5.

Messias, J., Schmidt, L., Oliveira, R., & Benevenuto, F. (2013). You followed my bot! transforming robots into influential users in twitter. *First Monday*, *18*(7).

Nix, A. (2016). *The power of big data and psychographics.* https://www.youtube.com/watch?v=n8Dd5aVXLCc.

Orwell, G. (1949). *1984.* Secker and Warburg.

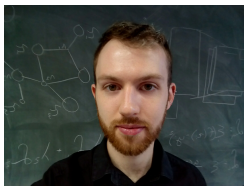
Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (n.d.). Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams. *Proceedings of the 20th International Conference Companion on World Wide Web*, 249–252.

Smith, D. V., & Delgado, M. R. (2015). Social nudges: utility conferred from others. *Nature Neuroscience*(6), 791–792.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., ... others (2016). Artificial intelligence and life in 2030. *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel*.

Zuckerberg, M. (2016). *Building global community.* https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10103508221158471/.

.



Dennis G Wilson is a PhD candidate at the Institut de Recherche en Informatique de Toulouse studying artificial neural models, after studying evolutionary computation and gene regulatory networks at the Anyscale Learning For All group at CSAIL, MIT. Their professional and personal information can be found at https://d9w.xyz/

RADIO AI - CALL FOR PARTICIPATION 2018

RADIO AI: An Important Invitation to AI Educators and Professionals to Talk About Your Work, Ideas and Thoughts On AI

Call for Participation: This invitation is extended across all the subfields of AI. The purpose of the RADIO AI project is to help educate the public and other professions about AI, with a crowd sourced collective of podcasts by people who work on AI. Submit podcasts in .mp3 or .wav by email to cmason@radioai.net See http://www.radioai.net for examples.

Some of the people talking the loudest about AI right now don't actually work on AI. We hope to hear from many individuals who wish to counter fear of AI by educating the public with lighthearted informative podcasts. People who are passionate about their work in AI tend to see the good it can do to help society, the environment, healthcare, and all aspects of life. We hope you will share this vision with others. Our goal is to bring together the collective vision through short podcasts and create easy to understand informative lectures by the people who work in AI - academia, business, finance, healthcare, inventors, and programmers. The lectures can be as short as 3 minutes or as long as 10. Some of you might wish to do a series of podcasts.

Deadlines:

Intent to submit: Nov. 1, 2017

Notifications Due: Dec. 1, 2017

Draft podcasts due: Dec. 12, 2017

Final podcasts due: Jan. 12, 2017

Audience: Please gage the audience as either teenage or in another field of study that is non-technical.

Directions: Example podcasts are located on www.radioai.net All you need is some peace and quiet, a microphone and your inspired thinking.

Topics:

History of AI, Software agents, Robots, Machine learning, Fuzzy AI, Overview of AI, AI and Society, Human-Robot Interaction and Applications - Healthcare, Legal, Transportation, Energy, Environment, etc. All topics related to AI are welcome. Other - if you're working in this field and are aware of the social changes that will or already have been taking place, we welcome your insights for the public.

Travel/hotel: There is no travel or hotel reservation required.

Contact: Dr. C. Mason cmason@radioai.net