How Artificial Intelligence Is Dividing Us Today, and How Terminator Might Kill Us Tomorrow

Hurford Center for the Arts and Humanities, Student Seminar Proposal

Keeton Martin and Harry Taussig

Faculty Advisor: Sorelle Freidler

17 March 2021

"The way to think about it is as 2.5 billion Truman Shows. Each person has their own reality with their own facts. Over time you have the false sense that everyone agrees with you because everyone in your news feed sounds just like you. Once you're in that state, it turns out you're easily manipulated."

- Roger McNamee

"How do you ethically steer the thoughts and actions of two billion people's minds every day?"

- Tristan Harris

Introduction

The dangers of developing advanced artificial intelligence are already upon us. Worldwide we are suffering from increased polarization, disinformation, information overload, and poor mental health exacerbated by the influence of AI. Furthermore, because algorithms are often trained on biased data, or written by biased people, the negative effects of AI are already disproportionately harming underprivileged groups. Although right now artificial intelligence can only slightly manipulate human behavior, it does so to large populations all at once through our smartphones and computers, causing huge impact. This impact will only continue to grow and some prominent scientists put the development of increasingly intelligent AI as the largest risk to the survival of future human generations.

We originally planned to split the seminar into two sections, where the first section would explore the current problems and risks already created by AI. Our current understanding is that societal problems caused by AI are already serious and complex, centering around the impact of individually cultivated streams of media and biased data. We each have our own personal stream of content curated for us by AIs that make it easy for us to get in personal echo chambers. Platforms that take advantage of this personal content delivery completely change our relationship with information and each other. The second section would be an exploration of the ways AI will pose existential risks to human life and values over the next few decades and beyond.

Our new plan for structuring the seminar is to lead discussions on current and future risks of AI development for the first half of the seminar, and then allow the interests of our group to guide the direction of discussion after providing this broader overview. The readings listed below will give students important background understanding of the risks related to AI and will foster curiosities to learn more about sub-topics like predictive policing or terminator-style world takeovers.

Understanding the extent to which AI is currently harming our society, and the ways in which we might safely create an AI that can understand and act in accordance with human values is a complex problem that needs a diverse set of perspectives. **Just because our focus is AI doesn't mean that this problem can or even should only be explored by computer science majors.** The problem is intertwined with almost every discipline at Haverford:

Political Science helps us question how we can use AI policy to protect human interests and fair treatment. It will also help to analyze past, present, and future attempts to abuse the power of highly intelligent algorithms by individuals.

Biology, Neuroscience, and Psychology majors can better understand the human mind, exploring how an AI's structure might mirror a real brain, and how an AI might exploit or take advantage of the way our minds are structured.

Philosophy and History majors can help us discuss what we actually should value, and how we could possibly encode that into the machines we are empowering. Would an equally intelligent or conscious machine deserve equal moral standing to a human? How have past technological revolutions impacted society, and how can we learn from our past mistakes and build off our previous knowledge?

English and Visual Studies helps us explore how we can tell compelling stories, and what stories we are currently telling ourselves about the intrusion of AI into our daily lives. What will it mean for us on a human, personal level to have our society overturned so quickly? How can we connect people with this pressing problem that feels so abstract?

Anthropology and Sociology skills will help us assess how likely we are to integrate and reject these new technologies. Will we be able to adapt to these new problems, or will they go over the general public's head? How will our societies and relationships change at the result of increasingly ever-present AI? How are our decisions affected when we increasingly trust AI's to make them for us? What about when they can make decisions for us better than we can ourselves?

Fundamental Questions:

What are the risks of developing advanced AI and Artificial General Intelligence?

What are the current symptoms we are experiencing from this problem?

How can we effectively regulate rapid innovations in AI?

What are the existential threats to humanity's long-term future caused by AI?

Definite Sources (first three weeks) -

- The Social Dilemma, Tristan Harris - Documentary

- The Coded Gaze, Rossi Films - Unpacking Biases in Algorithms That Perpetuate Inequity - Documentary
- AI Governance: Opportunity and Theory of Impact, Allan Dafoe
- Superintelligence FAQ, Scott Alexander

The final three weeks of content will be determined based on the discussion group's interests around AI risk.

Potential Speakers:

Marietje Schaake - A Dutch politician and served as a Member of the European Parliament from 2009-2019. She is a member of D66, part of the Alliance of Liberals and Democrats for Europe (ALDE) political group. She is Coordinator on the International Trade committee, where she is the ALDE spokesperson on transatlantic trade and digital trade. Schaake also serves on the committee on Foreign Affairs and the subcommittee on Human Rights. She is the Vice-President of the US Delegation and serves on the Iran Delegation and the Delegation for the Arab peninsula. Furthermore, Schaake is the founder of the European Parliament Intergroup on the Digital Agenda for Europe. In 2017 she was Chief of the European Union Election Observation Mission in Kenya. Since 2014, Schaake is a 'Young Global Leader' with the World Economic Forum and she was recently appointed as co-chair of the WEF Global Future Council on Agile Governance. Schaake is a Member of the Transatlantic Commission on Election Integrity, the Global Commission on the Stability of Cyberspace and chair of the CEPS Taskforce on Software Vulnerability Disclosure in Europe. Furthermore, she is a member of the European Council on Foreign Relations and an advisor to the Center for Humane Technology. Schaake was featured by Politico as one of the 28 most influential Europeans in the 'class of 2017'. Schaake is a HAI International Policy Fellow.

Yoshua Bengio - Recognized as one of the world's leading experts in artificial intelligence and a pioneer in deep learning.

Since 1993, he has been a professor in the Department of Computer Science and Operational Research at the Université de Montréal. CIFAR's Learning in Machines & Brains Program Co-Director, he is also the founder and scientific director of Mila, the Quebec Artificial Intelligence Institute, the world's largest university-based research group in deep learning.

In 2018, Yoshua Bengio ranked as the computer scientist with the most new citations worldwide, thanks to his many high-impact contributions.

In 2019, he received the ACM A.M. Turing Award, "the Nobel Prize of Computing", jointly with Geoffrey Hinton and Yann LeCun for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing.

Danah Boyd - A Partner Researcher at Microsoft Research, the founder and president of Data & Society, and a Visiting Professor at New York University. Her research is focused on addressing social and cultural inequities by understanding the relationship between technology and society. Her most recent books - "It's Complicated: The Social Lives of Networked Teens" and "Participatory Culture in a Networked Age" - examine the intersection of everyday practices and social media.

Susan Dumais - I am a technical fellow and managing director of Microsoft Research New England, Microsoft Research New York City and Microsoft Research Montreal. My research is at the intersection of information retrieval and human-computer interaction. I am interested in algorithms and interfaces for improved information retrieval, as well as general issues in human-computer interaction.

I have been at Microsoft Research since July 1997. My current research focuses on gaze-enhanced interaction, the temporal dynamics of information systems, user modeling and personalization, novel interfaces for interactive retrieval, and search evaluation. Previous research studied a variety of information access and management challenges, including personal information management, desktop search, question answering, text categorization, collaborative filtering, interfaces for improving search and navigation, and user/task modeling. I have worked closely with several Microsoft groups (Bing, Windows Desktop Search, SharePoint Portal Server and Office Online Help) on search-related innovations.