Loyola Marymount University

Ethical Artificial Intelligence

Mackenzie Drewe

In an age where technology is rapidly advancing and artificial intelligence is coming increasingly into focus, new questions are starting to arise. Researchers are trying to discover the best way for artificial intelligence to learn how to complete tasks with less help from humans. Currently, the most effective way for machines to learn is adaptive machine learning, which gives machines minimal instructions for a task and lets them learn on their own through repeated trials. It is unknown everything a machine could accomplish using this type of learning, including if they could learn how to be ethical. If it were possible for machines to learn ethics using adaptive machine learning, it would create unknown consequences for humans that are commonly explored in films and television. Combining the fictional portrayal of ethical machines in film with the reality of teaching machines ethics through adaptive machine learning, there is a lot to be uncovered before technology gets too ahead of us.

Mentor: Dr. Susan Scheibler

Discipline: Film Studies

Introduction

Many current experiments involving artificial intelligence (AI) are searching for the best ways to get machines to learn how to complete tasks on their own as opposed to being told how through programming by humans (Shead). One of the most recent and effective ways, adaptive machine learning, involves only giving a machine a goal and sometimes background information, while not providing how to specifically complete a task. The machine then learns on its own the most efficient way to accomplish said task through reinforcement learning, where the machine adapts through repeated trials (Shead).

Well known films and television shows have depicted artificial intelligence at the stage researchers eventually want to reach. They explore the consequences of incredibly advanced robots in the world interacting with humans. Some of the main issues presented in these films, or missing from these films, involve ethics; one issue being the ethics of the decisions the robots make (*Ex Machina; Her; 2001: A Space Odyssey; Battlestar Galactica*). Cinema can explore the idea of ethical machines in fictional scenarios, however it is unclear whether artificial intelligence in real life would ever be able to learn ethics. This is where adaptive machine learning and reinforcement learning tie in. Can artificial intelligence learn ethics through adaptive machine learning? If so, what would be the consequences resulting from ethical machines interacting with humans?

Background

Artificial intelligence has been discussed in philosophy and other realms of study for thousands of years, however the term was only coined in 1956 (Lewis). While there are years of artificial intelligence studies to focus on, Google's most recent experiments are current and center on adaptive machine learning. Google has an artificial intelligence division called

DeepMind that was created in 2014. Their mission is “to solve general intelligence and make machines capable of learning things for themselves” (Shead).

The main way DeepMind tests and trains their computer systems is by having them play Space Invaders, an old arcade game from the 1970s. DeepMind cofounder Mustafa Suleyman states, “Everything is learned completely from scratch - there’s absolutely zero pre-programmed knowledge so we don’t tell the agent these are Space Invaders or this is how you shoot” (Shead). Using reinforcement learning by having the systems play the game hundreds of times and only having the goal to optimize score, the systems were experts after only 500 games (Shead).

Another experiment from DeepMind involved a computer system called AlphaGo that was designed to understand and play the game of Go, a 3,000-year-old Chinese game more complicated than chess. The system was provided with a description of the board, shown a multitude of games between strong amateurs, and used reinforcement learning by playing the game against versions of itself thousands of times. Alpha Go became the first computer program to defeat a professional Go player (“The story of AlphaGo”).

DeepMind also created an AI that was simply programmed to get from one point to another. With only these instructions, visual sensors, and information about the objects around it, the AI learned how to walk, run, and jump with no other guidance (“Google’s DeepMind AI”). These three examples all demonstrate adaptive machine learning because each AI was not told how to play Space Invaders, play Go, or walk, but learned on their own through repetition and adaptation.

DeepMind has stated they are using these experiments to improve features of Google as well as learn about AI in general (Shead). They have not applied adaptive machine learning to everything they could and one field that has been debated but not experimented is ethics. Two

industries that rely on ethics to thrive are driverless cars and Carebots. Dr. Amy Rimmer, the lead engineer on the Jaguar Land Rover autonomous car, says the main issue keeping driverless cars off the roads is ethics, rather than any mechanical issues ("Can We Teach Robots Ethics?"). Carebots would be designed to help the sick and elderly and are expected to sprout in the next ten years ("Can We Teach Robots Ethics?"). Ethical machines are also explored in films and television shows, which provide interesting scenarios involving these machines living with humans. This research asks whether actual artificial intelligence could learn ethics through adaptive machine learning and would the consequences be similar to those portrayed in cinema.

Methods

Because I do not possess the knowledge or resources to conduct my own experiments, I plan to further study experiments done by DeepMind as well as any other companies researching adaptive machine learning. I will regularly visit DeepMind's website, deepmind.com, and look up news regarding artificial intelligence.

An abundance of published scholarly works has benefitted my research thus far and I will finish reading them and dive deeper into the sources I have already finished. One of those works is called *Common Sense, the Turing Test, and the Quest for Real AI* by Hector J. Levesque, which compares how humans and AI think and will be helpful in learning the technical process of how machines learn. Another source is *Robots are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws* by John Frank Weaver. This book looks at the realities of AI living with humans from a legal standpoint and is incredibly applicable to the second part of my research question that asks what the consequences would be of ethical machines interacting with humans. A third source, *Cylons in America: Critical Studies in Battlestar Galactica*, edited by Tiffany Potter and C. W. Marshall, analyzes a television show

that has to do with artificial intelligence, *Battlestar Galactica*, which lines up perfectly with my research.

While further investigating scholarly works that explore the first part of my question- can artificial intelligence learn ethics through adaptive machine learning- watching films and television addresses the second part of the question- the consequences of if the answer to the first part is yes. I have already watched films and television shows that confront AI and ethics, and the sources I will watch to further my research include *Ghost in the Shell* directed by Mamoru Oshii, *Blade Runner* directed by Ridley Scott, *Blade Runner 2049* directed by Denis Villeneuve, *The Terminator* directed by James Cameron, *Westworld* created by Jonathan Nolan and Lisa Joy, *A.I. Artificial Intelligence* directed by Steven Spielberg, and *I, Robot* directed by Alex Proyas.

I can read and watch these sources at anytime or place therefore making it easy to expand my knowledge. To find more sources I plan to visit the Margaret Herrick Library, which is the primary storage for works of the Academy of Motion Picture Arts and Sciences. I will be able to visit on Mondays, Tuesdays, Thursdays, and Fridays when it is open (Margaret Herrick Library). I will also visit other libraries when my class schedule permits.

Conferences about artificial intelligence are another informative research method I will employ. Two of the closer conferences that will take place in 2018 are the O'Reilly Artificial Intelligence Conference in San Francisco, California and the Singularity University Summit in San Diego, California (Zhou). I will also be able to watch conferences online or find articles summarizing the events.

Expected Results

The product of my research will be two fold. The first part will be an essay that teases out what I will learn through the methods described above. It will include the reality of whether AI

could learn ethics through adaptive machine learning and outline what others have already discovered. The parameters that would have to be taken if ethical machines joined humans will also be considered as well as possible results and consequences of this interaction. The second component will be a screenplay. Watching films and television helps in the creation of this product because it shows what has already been portrayed in media and how I can make my work different. The screenplay will allow me to craft scholarly questions in a creative way and will demonstrate my findings in the world of the story, either by creating a world in which artificial intelligence can learn ethics or one in which they cannot.

Conclusion

Researchers are discovering how artificial intelligence can learn more efficiently using adaptive machine learning. Films and television commonly depict ethical AI in fictional worlds. Can machines learn ethics through adaptive machine learning? If so, what would be the consequences resulting from ethical machines interacting with humans? Studying scholarly works and attending conferences will provide information on the logistics of if AI could actually learn ethics using adaptive machine learning. Watching and analyzing films and television provides possible consequences of ethical machines living with humans in fictional worlds. Through my first deliverable, an essay, I will add to current scholarly works that discuss adaptive machine learning. My second deliverable, a screenplay, will join the branch of the film industry focused on artificial intelligence.

Works Cited

A.I. Artificial Intelligence. Directed by Steven Spielberg, performances by Haley Joel Osment and Jude Law, Warner Bros. Pictures, 2001.

Blade Runner. Directed by Ridley Scott, performances by Harrison Ford, Rutger Hauer, Sean Young, Edward James Olmos, Warner Bros. Pictures, 1982.

Blade Runner 2049. Directed by Denis Villeneuve, performances by Ryan Gosling, Harrison Ford, Ana de Armas, Warner Bros. Pictures, 2017.

"Can We Teach Robots Ethics?" *BBC News*, 15 Oct. 2017, www.bbc.com/news/magazine-41504285. Date Accessed 15 Oct. 2017.

Ex Machina. Directed by Alex Garland, performances by Alicia Vikander, Oscar Isaac, Domhnall Gleeson, Universal Studios, 2015.

Ghost in the Shell. Directed by Mamoru Oshii, performances by Atsuko Tanaka, Akio Ōtsuka, and Iemasa Kayumi, Shochiku, 1995.

"Google's DeepMind AI just taught itself to walk." *YouTube*, uploaded by Tech Insider, 12 July 2017, https://www.youtube.com/watch?v=gn4nRCC9TwQ. Date Accessed 21 Nov. 2017.

Her. Directed by Spike Jonze, performances by Joaquin Phoenix, Scarlett Johansson, Amy Adams, Warner Bros, 2013.

I, Robot. Directed by Alex Proyas, performances by Will Smith, Bridget Moynahan, and Bruce Greenwood, 20th Century Fox, 2004.

Larson, Glen, creator. *Battlestar Galactica*. David Eick Productions, 2004.

Levesque, Hector J. *Common Sense, The Turing Test, and the Quest for Real AI*. London, The MIT Press, 2017.

Lewis, Tanya. "A Brief History of Artificial Intelligence." *Live Science*, Purch, 4 Dec. 2014, https://www.livescience.com/49007-history-of-artificial-intelligence.html. Date Accessed 9 Dec. 2017.

Li, Deyi, and Yi Du. *Artificial Intelligence with Uncertainty*. Boca Raton, CRC Press, 2017.

Margaret Herrick Library. Academy of Motion Picture Arts and Sciences, 2015, http://www.oscars.org/library. Date Accessed 9 Dec. 2017.

Nolan, Jonathan and Lisa Joy, creators. *Westworld*. Warner Bros. Television Distribution, 2016.

Potter, Tiffany, and C. W. Marshall. *Cylons in America: Critical Studies in Battlestar Galactica*. New York, The Continuum International Publishing Group Inc, 2008.

Shead, Sam. "Google DeepMind: What is it, how does it work and should you be scared?" *Techworld*, IDG UK, 15 Mar. 2016, https://www.techworld.com/apps-wearables/google-deepmind-what-is-it-how-it-works-should-you-be-scared-3615354/. Date Accessed 4 Dec. 2017.

Snyder, Larry. "Reinforcement Learning Explained Using the Beer Game." *SupplyChainDigest*, 1 Nov. 2017, **http://www.scdigest.com/experts/DrWatson_** 17-11-01.php?cid=13234. Date Accessed 28 Nov. 2017.

Solon, Olivia. "Deus Ex Machina: Former Google Engineer is Developing an AI God." *The Guardian*, 28 Sept. 2017, **www.theguardian.com/technology/2017/sep/28/artificial-intelligence-god-anthony-levandowski?CMP=Share_iOSApp_Other.** Date Accessed 13 Oct. 2017.

The Matrix. Directed by the Wachowski Brothers, performances by Keanu Reeves, Laurence Fishburne, and Carrie-Anne Moss, Village Roadshow Pictures, 1999.

"The story of AlphaGo so far." *Deep Mind*, DeepMind Technologies Limited, 2017, **https://deepmind.com/research/alphago/**. Date Accessed 28 Nov. 2017.

The Terminator. Directed by James Cameron, performances by Arnold Schwarzenegger, Michael Biehn, Linda Hamilton, and Paul Winfield, Orion Pictures, 1984.

2001: A Space Odyssey. Directed by Stanley Kubrick, performances by Keir Dullea, Gary Lockwood, Douglas Rain, Metro-Goldwyn-Mayer, 1968.

Weaver, John Frank. *Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws*. Santa Barbara, Praeger, 2014.

Zhou, Adelyn. "Best Artificial Intelligence Conferences For 2018." *TOPBOTS*, 7 Nov. 2017, www.topbots.com/best-artificial-intelligence-conferences-2018/. Date Accessed 9 Dec. 2017.

Budget

My budget will consist of paying for books, films, television shows, transport to libraries and conferences, and motel rooms for attending conferences not in Los Angeles. Books can range anywhere from $7.00-$30.00 and films and television shows can cost around $20.00. Assuming I read ten books, I will need $200.00 to buy or rent them. Knowing I will watch seven films/television shows, I will need $140.00. Transportation should cost about $100.00 if I stay close to Los Angeles and do not include flights. Cheaper motel rooms cost around $75.00. Assuming I stay at two motels I will need $150.00. Adding these totals up equals $590.00. To be safe, I would like to request $700.00 to fund my research.