Ethical Implications Of Bias In Machine Learning

Adrienne Yapo
Bentley University
175 Forest St.
Waltham, MA 02452
adrienne.yapo@gmail.com

Joseph Weiss
Bentley University
175 Forest St.
Waltham, MA 02452
jweiss@bentley.edu

Abstract

Biases in AI and machine learning algorithms are presented and analyzed through two issues management frameworks with the aim of showing how ethical problems and dilemmas can evolve. While "the singularity" concept in AI is presently more predictive than actual, both benefits and damage that can result by failure to consider biases in the design and development of AI. Inclusivity and stakeholder awareness regarding potential ethical risks and issues need to be identified during the design of AI algorithms to ensure that the most vulnerable in societies are protected from harm.

1. Introduction

The concept of Artificial Intelligence (AI) stretches to Greek antiquity with the automatons of the blacksmith Hephaestus — the Hounds of Alcinous, the Tripods of Olympus, the Keledones of Apollo — self-operating creatures crafted of metal to carry out the work of the Gods ("Hephaestus, god of craftsmanship, blacksmiths and stonemasonry — Greek gods, mythology of ancient Greece," 2013). Indeed, a common perception of AI is that of smart robot droids operating alongside humans. Benevolent ideals of AI in film, such as *Star Wars* C-3P0 & R2D2 and *Wall-E*, reflect the hope, dreams, and anticipation of AI. However, just as many, if not more, relate cautionary and dystopian tales of AI run amok. Consider the silent film, *Metropolis* (1927), about a deadly "female" robot built by a grieving mad scientist, or the malevolent computer "Hal" who kills most of the crew in *2001: A Space Odyssey* (1968). *War Games* (1983) put the world on the brink of nuclear disaster as a military computer system tries to "win" the game of war by nearly starting World War III. *The Matrix* (1999) imagines an evil "singularity" (when machines attain the ability to surpass human intelligence) that enslaves the human race behind a fake veil of utopian reality [35].

In actuality, the field of AI has experienced several starts and stops since its official beginning in 1956 [19], but while predictions of "the singularity" have not as yet come to pass. AI is currently experiencing a period of exponential growth. From self-driving cars and lethal autonomous weapons systems to machine learning algorithms and AI powered IoT devices, AI is rapidly transforming society and our world in unprecedented ways. The accelerated advances in AI, especially related to machine learning algorithms, are having far-reaching and profound consequences on our lives. Ethical considerations for consumers, society, public policy, laws, and regulation are beginning to form.

While several scholars have discussed the potential risks of AI [9; 46] and ethical issues [7, 22, 43], our contribution presents two issues management frameworks within which to view and identify different ethical issues, which can also serve to alert stakeholders to potential moral harms that AI endeavors may present.

1.1. The rise of machine learning

Machine learning is a subset of AI where algorithms directed by complex neural networks teach computers to think like a human while processing "big data" and calculations with high precision, speed, and supposed *lack of bias* [28].

"The development of neural networks has been key to teaching computers to think and understand the world in the way we do, while retaining the innate advantages they hold over us such as speed, accuracy and lack of bias. A Neural Network is a computer system designed to work by classifying information in the same way a human brain does. It can be taught to

URI: http://hdl.handle.net/10125/50557
ISBN: 978-0-9981331-1-9

recognize, for example, images, and classify them according to elements they contain. Essentially it works on a system of probability — based on data fed to it, it is able to make statements, decisions or predictions with a degree of certainty. The addition of a feedback loop enables 'learning' — by sensing or being told whether its decisions are right or wrong, it modifies the approach it takes in the future" [28].

Machine learning algorithms are already interwoven into many aspects of our daily lives, influencing in unseen ways as decisions are made for us and about us with little to no transparency. Web search and recommendation machine learning algorithms drive relevant search results and product recommendations from the likes of Google, Netflix, and Amazon. Facebook's facial recognition uses a machine learning algorithm to automatically identify and tag friends when uploading a photo [15] and the news feed algorithm prioritizes posts for what it thinks we'd most like to see. Machine learning in medicine is providing important advances in health care and treatment decisions while AI in computational biology / drug discovery is increasing the pace of research [19]. The Finance industry utilizes machine learning algorithms to uncover credit card fraud, make predictions about creditworthiness, and identify trends in the stock market. And the criminal justice system is using machine learning to predict crime hotspots and recidivism rates [1].

1.2. Machine learning algorithm bias

Although machine learning algorithms can produce numerous benefits to individuals, consumers, businesses, investors, the government, and society at large, recent research has uncovered many instances of bias in machine learning algorithms that have troubling implications and deleterious consequences.

In 2015, academia and the media sources reported instances of apparent gender bias in Google search. Top results for a "CEO" image search returned only photos of white men. Shortly thereafter, a research study at Carnegie Mellon University revealed that Google displayed significantly less ads for high-paying executive jobs if the search engine thought a female job-seeker was conducting the search. The team found that "Google shows the [high-paying executive job] ads 1,852 times to the male group — but just 318 times to the female group." Professor Annupam Datta was quoted on the study: "I think our findings suggest that there are parts of the ad ecosystem where kinds of discrimination are beginning to emerge and there is lack of transparency. This is concerning from a societal standpoint." [11] He adds, "Many important decisions in society these days are being made by algorithms. These algorithms run inside of boxes we don't have access to the internal details of. The genesis of this project was that we wanted to peek inside this box a little to see if there are more undesirable consequences of this activity going on."

Similarly, racist algorithms have recently been identified in digital photo technology. In May of 2015, Flickr's image recognition tool was reportedly displaying racist results, tagging black people as "animals' or "apes." Hewlett-Packard's software for web cameras struggled to recognize dark skin tones, and Nikon's camera software was inaccurately identifying Asian people as blinking. In Kate Crawford's 2016 *New York Times* article "Artificial Intelligence's White Guy Problem," she blamed the bias on bad data and lack of inclusivity. "Algorithms learn by being fed certain images, often chosen by engineers, and the system builds a model of the world based on those images. If a system is trained on photos of people who are overwhelmingly white, it will have a harder time recognizing nonwhite faces." [14]

More recently, the *New York Times* reported that Facebook executives, the morning after the 2016 Presidential election, internally considered what influence the company had played in the election results. Public accusations from multiple sources mounted that its news feed algorithm had distributed deliberate fake news stories and misinformation that unfairly biased voters against Hillary Clinton and influenced the election outcome in favor of Donald Trump, while increasing societal divisions among Americans through "filter bubbles." [21] While the examples of racial and gender bias in Google Search and digital photo technology were more covert, Facebook's algorithm regularly distributes overtly racist, sexist, and alt-right content to billions, normalizing such content through its distribution on legitimate and mainstream platforms.

1.3. Machine learning in the criminal justice system

Yet perhaps the most troubling incidents of bias in machine learning to date are unfolding in the criminal justice system. Consider the following statement from then U.S. Attorney General Eric Holder on the Sentencing Reform and Corrections Act of 2015:

"Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice...they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system, and in our society."

The Sentencing Reform and Corrections Act of 2015 (Bill S.2123), still pending in Congress, seeks to include the implementation of mandatory risk and needs assessment systems in all federal prisons. These systems are used to evaluate recidivism rates and assign scores indicating whether a particular defendant is low, medium, or high risk to commit future crimes. Risk assessments are being used throughout all phases of America's criminal justice system, from pre-trial release to parole and everything in between. In nine states (Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington, and Wisconsin), these scores are provided to judges for consideration during sentencing (Angwin, Larson, Mattu, and Kirchner, 2016) [1].

Risk assessment systems are driven by complicated machine learning algorithms that calculate scores based on a number of variables including employment history, education levels, and prior crimes, among others. In fact, risk assessment tools were designed with the goal of overcoming judicial and sentencing bias. The machine learning algorithms that assess risk in pre-trial release and recidivism rates expressly do not include race and ethnicity in their calculations, so theoretically, they should produce results that are unbiased and objective.

Attorney General Holder was prescient to express concern about "exacerbating unwarranted and unjust disparities." According to an in-depth investigation by *ProPublica* (Angwin, Larson, Mattu, and Kirchner, 2016) [1], COMPAS, a popular for-profit risk assessment product developed by Northpointe and used by judicial systems throughout the U.S., is twice as likely to mistakenly identify white defendants as a low risk for committing future crimes, and twice as likely to erroneously tag black defendants as a high risk. Consider these results from ProPublica's study:

Table 1: Disproportionate incarceration rates

Source: ***Propublica*** **analysis from Broward County, Florida**

Prediction Fails Differently for Black Defendants

	White	African American
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

"Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes."

So how can an algorithm that doesn't even factor race and ethnicity into its calculation produce such biased results? A significant issue with bias in machine learning algorithms is the "black-box" secrecy behind their design. For-profit companies that produce these algorithms do not release the criteria and calculations behind the formulas. The algorithms are also often so complex that even the engineers and designers that have access to the formulas may struggle or fail to predict the outcome and effects of the algorithms results. As such, because these algorithms are created by humans, they inevitably — and often unconsciously — reflect societal values, biases, and discriminatory practices (Centre for Internet and Human Rights, 2017).

The US has one of the highest incarceration rates in the world (Ye Hee Lee, 2015) [43] and black people are disproportionately affected, as Table 1 indicates. According to the NAACP "Criminal Justice Fact Sheet," black people in the U.S. are incarcerated at close to six times the rate of white people. Moreover, in 2008 black and Hispanic people accounted for 58% of the prison population while only representing 25% of the total population. It is clear that deep societal and systemic biases against minorities exist in the U.S. A recent follow-up story from *ProPublica*, "Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say," released on December 30, 2016, reported on results of four independent studies that explored whether a formula could be created "that is equally predictive for all races without disparities in who suffers the harm of incorrect predictions." Significantly, they found that these risk assessment algorithms "have been written in a way that guarantees black defendants will be inaccurately identified as future criminals more often than their white counterparts."

"It's actually impossible for a risk score to

satisfy both fairness criteria at the same time. The problem, several said in interviews, arises from the characteristic that criminologists have used as the cornerstone for creating fair algorithms, which is that formula must generate equally accurate forecasts for all racial groups. The researchers found that an algorithm crafted to achieve that goal, known as 'predictive parity,' inevitably leads to disparities in what sorts of people are incorrectly classified as high risk when two groups have different arrest rates.

"Predictive parity' actually corresponds to 'optimal discrimination,'" said Nathan Srebro, associate professor of computer science at the University of Chicago and the Toyota Technological Institute at Chicago. That's because predictive parity results in a higher proportion of black defendants being wrongly rated as high-risk."

However, the article does suggest that these algorithms could be revised to introduce fairness in the algorithm to account for the inherent underlying societal bias of higher arrest rates for black people. Unfortunately, Broward County officials in Florida — where the *ProPublica* study was conducted — has not changed any policies with regard to utilizing COMPAS data in light of the recent findings. Further investigation and research studies will perhaps help inform future design decisions in risk assessment algorithms and public policy decisions regarding the use of risk assessment systems in criminal justice.

2. Ethics of algorithms – the way forward

As the ethical implications and consequences of AI have already begun to affect our lives, the development of ethics, standards, and regulation considerations for AI is emerging as important decisions in the technology sector. Utilizing Fink's Seven-Phase Issue-Development Process framework [17; 42, p. 144] we present the evolution of AI issues that relate to the public at a (U.S.) societal level, particularly with regard to machine learning algorithms with the aim of anticipating and preventing future harm to the public.

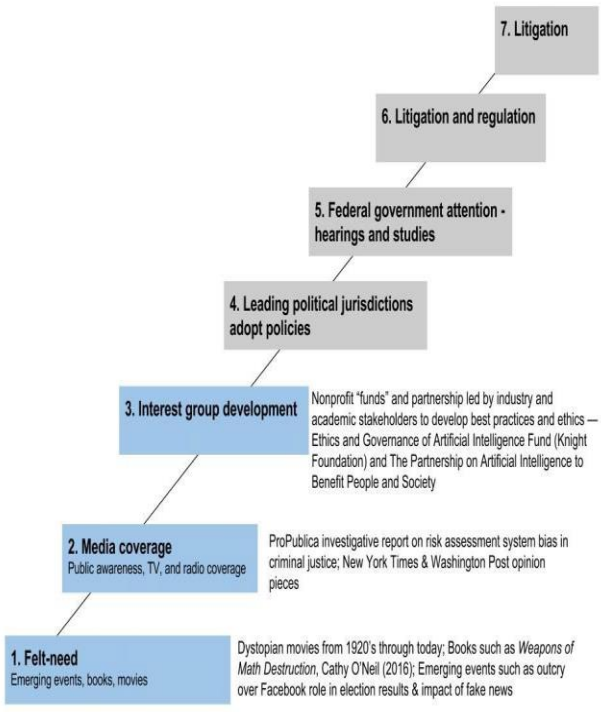

Figure 1: Marx's 7 Stages of Issue Development

Source: Adapted from Marx, T. (1986). Integrating public affairs and strategic planning. California Management Review, Fall, 29(1), 145.

Figure 1 illustrates the first phase of Mark's [26;] model of issues development, "Felt-need," which is triggered by any number of sources such as emerging events, books, movies, and precipitating crises. We propose that recently, non-fiction books, public discourse, and actual events have underscored the potential harm that AI can cause. Cathy O'Neil's 2016 [31] book *Weapons of Math Destruction* explores the harmful consequences of "opaque" and "unregulated" big data mathematical models that reinforce inequality and discrimination.

The outcry in public discourse on Facebook's potential role in disseminating fake news and influencing election outcomes, partisanship and divisiveness in our society is being examined by various stakeholders, including individuals, technology companies, and sectors of the government.

Phase two, "Media coverage," is evident in the multitude of news stories and opinion pieces produced on a daily basis. One of the most

important pieces of media so far on this issue, as outlined in Part I, is the *ProPublica* investigative reports, and the studies undertaken by independent researchers as a result of these reports, on bias in the risk assessment algorithms that are being widely used in the criminal justice system. Likewise, Kate Crawford's 2016 *New York Times* opinion piece implored vigilance in machine learning design:

"Sexism, racism, and other forms of discrimination are being built into the machine-learning algorithms that underlie technology behind many 'intelligent' systems that shape how we are categorized and advertized to...We need to be vigilant about how we design and train these machine-learning systems, or we will see ingrained forms of bias built into the artificial intelligence of the future.

Like all technologies before it, artificial intelligence will reflect the values of its creators. So inclusivity matters — from who designs it to who sits on the company boards and which ethical perspectives are included. Otherwise, we risk constructing machine intelligence that mirrors a narrow and privileged vision of society, with it sold, familiar biases and stereotypes" [14]

We argue that this process has not as yet evolved farther than Phase Three — "Interest Group Development"—which is starting to take shape, although businesses, governmental institutions, and public awareness of bias in AI machine learning algorithms is growing; policies at the city, county, state or federal level have not changed as yet evolved.

In September 2016, *The Partnership on Artificial Intelligence to Benefit People and Society* was announced. This ten-member board with representatives from Amazon, Facebook, Google, IBM, and Microsoft will "conduct research and recommend best practices relating to 'ethics, fairness and inclusivity; transparency, privacy, and interoperability; collaboration between people and AI systems; and the trustworthiness, reliability and robustness of the technology.' It does not intend to lobby government or other policymaking bodies." (Parloff, 2016) [32].

In early January 2017, another initiative was announced, *The Ethics and Governance of Artificial Intelligence Fund*, which is a $27 million fund that has been established to ensure ethical behavior and governance in developing AI technologies. Partners include the John S. and James L. Knight Foundation, Omidyar Network, LinkedIn founder Reid Hoffman, The MIT Media Lab, and the Berkman Klein Center for Internet & Society at Harvard University. According to the press release, the fund will address issues such as:

- *Communicating complexity: How do we best communicate, through words and processes, the nuances of a complex field like AI?*
- *Ethical design: How do we build and design technologies that consider ethical frameworks and moral values as central features of technological innovation?*
- *Advancing accountable and fair AI: What kinds of controls do we need to minimize AI's potential harm to society and maximize its benefits?*
- *Innovation in the public interest: How do we maintain the ability of engineers and entrepreneurs to innovate, create and profit, while ensuring that society is informed and that the work integrates public interest perspectives?*
- *Expanding the table: How do we grow the field to ensure that a range of constituencies are involved with building the tools and analyzing social impact? (Knight Foundation, 2017)* [23].

Another complementary approach to the above framework, is Fink's [17, 42] four stages of crisis management, which we also use to analyze the ethical implications of AI, specifically with regard to bias in machine learning algorithms. This model has a "precrisis" phase designed to signal warning signs and symptoms of an issue if and/or before it could become a crisis.

We observe that AI's effects on the U.S. society is at Stage 1 — the "precrisis" or "prodromal stage," since there are many warnings and symptoms occurring as evidenced by media reports and research studies, several of which were presented earlier. The establishment of the two AI Ethics partnerships is also evidence that technology companies and academic research institutions are recognizing the warning signs and will hopefully work to prevent Stage 2 of this model — "the Acute Crisis Stage / Point of No Return" — from occurring.

Taken together, the two issues management approaches summarized above are summarized here to alert active participants and relevant stakeholders and stockholders in AI fields to potential ethical dilemmas and harm, as well as positive contributions, that AI transformations through products and services may present.

Lastly, the role of Corporate Social Responsibility and the so-called "Carrot," values-based, vs. "Stick," rules-based, approaches to stakeholder management are relevant to the development and implementation of AI technologies [42]. Technology companies are establishing voluntary self-regulation —a standard values, ethics, best practices, risk management, and philanthropy — to prevent the government from imposing external regulation and congressional oversight. The establishment of *The Partnership on Artificial Intelligence to Benefit People and Society* and *The Ethics and Governance of Artificial Intelligence Fund* indicates direct evidence of this. Whether or not and the extent to which local, state, and federal legislation will emerge in response to different stakeholders' interests, rights, and responsibilities is yet to be seen.

3. Conclusion

The Ethics and Governance of Artificial Intelligence Fund issued an original press release the Knight Foundation stating: "Because of this pervasive but often concealed impact, it is imperative that AI research and development be shaped by a broad range of voices—not only by engineers and corporations, but also by social scientists, ethicists, philosophers, faith leaders, economists, lawyers and policymakers."[23]

While "the singularity" concept in AI is presently more predictive than actual, both benefits and damage that can be caused by failure to consider bias in the design and development of AI is present. As has been illustrated by the *ProPublica* investigation and other examples offered here, inclusivity and stakeholder awareness of impending ethical risks and issues are crucial in the design of AI to ensure that the most vulnerable in our society are protected from harm. If self-regulation alone fails to prevent and correct many negative biases in AI design processes sponsored by corporations, compliance through legislative and enforcement of standards may likely result.

Bibliography

[1] Angwin, J., & Larson, J. (2016, December 30). Bias in criminal risk scores is mathematically inevitable, researchers say. Retrieved February 1, 2017, from Machine Bias, https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-rese archers-say

[2] Angwin, J., Mattu, S., Larson, J., & Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. Retrieved January 22, 2017, from Machine Bias, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Ashrafian, H. (2015). AIonAI: A humanitarian law of artificial intelligence and robotics. Science and engineering ethics, 21(1), 29-40.

[4] Ashrafian, H. (2015). Artificial intelligence and robot responsibilities: Innovating beyond rights. Science and engineering ethics, 21(2), 317-326.

[5] Baldridge, J. (2015, August 2). Machine learning and human bias: An uneasy pair. Retrieved January 29, 2017, from Tech Crunch, https://techcrunch.com/2015/08/02/machine-learning-and-human-bias-an-uneasy-pair/

[6] Basile, D. (2017, January 22). Ethics — the next frontier for artificial intelligence. Retrieved January 22, 2017, from https://techcrunch.com/2017/01/22/ethics-the-next-frontier-for-artificial-intelligence/

[7] Beauchamp, T. and Chilress, J. (2001). Principles of Biomedical Ethics. Oxford: Oxford University Press.

[8]7] Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. The Cambridge Handbook of Artificial Intelligence, 316-334.

[9] Bostrom, N. (2002). 'Existential Risks: Analyzing Human Extinction Scenarios', Journal of Evolution and Technology 9.

10[8] Buttarelli, G. (2016, December 20). A smart approach: Counteract the bias in artificial intelligence. Retrieved January 29, 2017, from European Data Protection Supervisor, https://secure.edps.europa.eu/EDPSWEB/edps/pid/696

11 Carpenter, J. (2015, July 7). Google's algorithm shows prestigious job ads to men, but not to women. The Independent - News. Retrieved from http://www.independent.co.uk/life-style/gadgets-and-tech/news/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-0372166.html

12 Čerka, P., Grigienė, J., & Sirbikytė, G. (2015). Liability for damages caused by artificial intelligence. Computer Law & Security Review, 31(3), 376-389.

13 Coca, N. (2016, January 31). Why your favorite apps are designed to addict you. Retrieved January 22, 2017, from http://kernelmag.dailydot.com/issue-sections/features-issue-sections/15708/addicting-apps-mobil e-technology-health/

14 Crawford, K. (2016, December 5). Artificial intelligence's white Guy Problem. SundayReview. Retrieved from https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem. html

15 De Vos, T. (2016, December 5). Cool machine learning examples in real life. Retrieved February 5, 2017, from Algorithms, http://itenterprise.co.uk/cool-machine-learning-examples-real-life/

16 Devlin, H. (2017, February 2). Discrimination by algorithm: Scientists devise test to detect AI bias. The Guardian. Retrieved from https://www.theguardian.com/technology/2016/dec/19/discrimination-by-algorithm-scientists-de vise-test-to-detect-ai-bias

17 Fink, S. (1986). Crisis management: Planning for the inevitable. New York: American Management Association.

18 Hephaestus, god of craftsmanship, blacksmiths and stonemasonry - Greek gods, mythology of ancient Greece. (2013). Retrieved January 26, 2017, from Greek Gods, http://www.greek-gods.org/olympian-gods/hephaestus.php

19 Hamet, P., & Tremblay, J. (2017). Artificial Intelligence in Medicine. Metabolism.

20 Harvard President & Fellows (2017, January 20). Berkman Klein center and MIT media lab to collaborate on the ethics and governance of artificial intelligence - Harvard law today. Retrieved January 22, 2017, from http://today.law.harvard.edu/berkman-klein-center-mit-media-lab-collaborate-ethics-governance-artificial-intelligence/?utm_source=twitter&utm_medium=social&utm_campaign=hu-twitter-gen eral

21 Isaac, M. (2016, November 22). Facebook, in cross hairs after election, is said to question its influence. Technology. Retrieved from https://www.nytimes.com/2016/11/14/technology/facebook-is-said-to-question-its-influence-in-e lection.html?_r=0

22 Kamm, F. (2007). Intricate Ethics: Rights, Responsibilities, and Permissible Harm. Oxford: Oxford University Press.

23 Knight Foundation, Omidyar Network and Linkedin Founder Reid Hoffman Create $27 Million Fund to Research Artificial Intelligence for the Public Interest. Retrieved January 21, 2017, From Knight Foundation, http://www.knightfoundation.org/press/releases/knight-foundation-omidyar-network-and-linkedin-founder-reid-hoffman-create-27-million-fund-to-research-artificial-intelligence-for-the-public-interest

24 Larson, S. (2016, July 15). Research shows gender bias in Google's voice recognition. Retrieved January 29, 2017, from Debug, http://www.dailydot.com/debug/google-voice-recognition-gender-bias/ In-line Citation: (Larson, 2016)

25 Loukides, M. (2016, November 14). Mike Loukides. Retrieved January 22, 2017, from https://www.oreilly.com/ideas/the-ethics-of-artificial-intelligence

26 Marx, T. (1986), Integrating public affairs and strategic planning. California Management Review, Fall 29(1), 145.

27 Markoff, J. (2016, September 2). How tech giants are devising real ethics for artificial intelligence. Technology. Retrieved from https://www.nytimes.com/2016/09/02/technology/artificial-intelligence-ethics.html?_r=0

28 Marr, B. (2016, December 6). What is the difference between artificial intelligence and machine learning? Forbes. Retrieved from http://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/#3a953a62687c

29 NAACP. (2017). Criminal justice fact sheet. Retrieved February 1, 2017, from http://www.naacp.org/criminal-justice-fact-sheet/

30 Of Prediction and Policy. (2016, August 01). Retrieved January 29, 2017, from The Economist, http://www.economist.com/news/finance-and-economics/21705329-governments-have-much-gai n-applying-algorithms-public-policy

31 O'Neil, C. (2016). Weapons of Math Destruction. New York, New York: Crown.

32 Parloff, R. (2016, September 28). AI Partnership Launched by Amazon, Facebook, Google, IBM, and Microsoft. Retrieved February 1, 2017, from Fortune, http://fortune.com/2016/09/28/ai-partnership-facebook-google-amazon/

33 Pistono, F., & Yampolskiy, R. V. (2016). Unethical Research: How to Create a Malevolent Artificial Intelligence. arXiv preprint arXiv:1605.02817.

34 Reese, H. (2016, November 18). Bias in machine learning, and how to stop it. Retrieved January 29, 2017, from Tech Republic, http://www.techrepublic.com/article/bias-in-machine-learning-and-how-to-stop-it/ In-line Citation: (Reese, 2016)

35 Reid, J. (2016, February 19). 25 of the best movies about AI, ranked. Retrieved February 4, 2017, from ZD Net, http://www.zdnet.com/pictures/15-of-the-best-movies-about-ai-ranked/ Regular. (2017). How to make sure the future of AI is ethical. Retrieved January 22, 2017, from https://www.weforum.org/agenda/2016/12/lets-get-ethical-the-future-of-ai/Regular. (2017). Top 9 ethical issues in artificial intelligence. Retrieved January 22, 2017, from https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/

36 Revell, T. (2016, December 1). Concerns as face recognition tech used to "identify" criminals. Retrieved January 22, 2017, concerns-as-face-recognition-tech-used-to-identif y-criminals

https://www.newscientist.com/article/2114900-concerns-as-face-recognition-tech-used-to-identif y-criminals

37 Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. AI Magazine, 36(4), 105-114.

38 Russell, S. (2015). Ethics of artificial intelligence
.
39 Scherer, M. U. (2015). Regulating artificial intelligence systems: risks, challenges, competencies, and strategies.
Singer, A. E. (2015). Corporate moral agency and artificial intelligence. In Human Rights and Ethics: Concepts, Methodologies, Tools, and Applications (pp. 505-517). IGI Global.

40 Smith IV, J. (2016, October 11). Crime-prediction tool may be reinforcing discriminatory policing. Retrieved January 29, 2017, from Business Insider UK, http://uk.businessinsider.com/predictive-policing-discriminatory-police-crime-2016-10?r=US&I R=T

41 Vieth, K., & Bronowicka, J. Ethics of Algorithms. Retrieved January 29, 2017, from Centre for Internet and Human Rights, https://cihr.eu/eoa2015web/

42 Weiss, J. W. (2014). Business Ethics, A Stakeholder and Issues Management Approach. Barrett-Koehler Publishers, Oakland, CA, bkpub@bkpub.com

43 Wendell, W. (2008). 'Moral Machines: Teaching Robots Right from Wrong.' Oxford University Press.

44 Will Smart machines be less biased than humans? (2016). Retrieved January 29, 2017, from http://gereportsasean.com/post/153029327230/will-smart-machines-be-less-biased-than-humans

45 Ye Hee Lee, M. (2015, July 7). Yes, U.S. Locks people up at a higher rate than any other country. Washington Post. Retrieved from
https://www.washingtonpost.com/news/fact-checker/wp/2015/07/07/yes-u-s-locks-people-up-at- a-higher-rate-than-any-other-country/?utm_term=.bf4bf2800af1

46 Yudkowsky, E. (2008a.) Artificial Intelligence as a Positive and Negative Factor in Global Risk, in Bostrom and Cirkovic (eds.), pp. 308-345; and, (2008b). Cognitive biases potentially affecting judgment of global risks',
in Bostrom and Cirkovic (eds.), pp. 91-119.