References

Alaieri, F., & Vellino, A. (2016). Ethical decision making in robots: Autonomy, trust and responsibility. In *International conference on social robotics* (pp. 159–168).

Vellino and Alaieri discuss the ways in which robots as ethical agents could raise uncertainty and mistrust among humans due to their autonomy and decision making methods. Thus, the authors propose two approaches to assess the accountability of the actions committed by these robots. The 'top-down' approach allows the programmer to install decision making and considerably ethical behaviour algorithms; while the 'bottom-up' approach enables the robot to collect information from its surroundings and learn from its experiences. The first holds the programmer accountable while the latter, the robot. Vellino and Alaieri suggest a hybrid strategy, combining both approaches to allow ethical robots to have well-defined rules guiding their actions, while also allowing them to learn new ethical principles. Through these approaches, programmers can be regulated to implement the hybrid strategy or otherwise be made aware of the consequences of using one in favour of the other.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*.

This is a high level paper which discusses many of the potential safety problems with developing general artificial intelligence. As the complexity of AI systems increase it is inevitable that they will operate in more and more complex environments. Complex environments mean that it will be impossible to consider every possible action an AI may take even if it is extensively tested. Complex environments may also prove challenging or impossible to simulate, making it hard to know the effectiveness of training sessions. We plan on using this paper to discuss how negligent AI development may lead to catastrophic results.

Balahur, A., Mihalcea, R., & Montoyo, A. (2014). Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. *Computer Speech & Language*, *28*(1), 1–6. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0885230813000697` doi: https://doi.org/10.1016/j.csl.2013.09.003

This article gives an introduction on other papers that discuss methods and usages of sentiment analysis in a broad variety of situations. It demonstrates the versatility of this technology and how it might be used in the future. In relation to the workplace, this technology can be applied to screen and monitor employees at a previously unimaginable level, which can have severe ethical implications. We plan on using this paper to discuss how such a technology could impact the workplace.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349–4357).

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and information technology*, *15*(3), 209–227.

In this paper Bozdag discusses the ways in which the information provided by computers is dictated by algorithms susceptible to partiality. He focusses on how search engines and interactive technology is actively filtering the content humans access. The filtering consists of personalization algorithms, manager impact, and conscious selection of information by the companies in charge. This 'gatekeeping' of information may have unfavorable consequences but can not be applied universally. Nonetheless, biases in data acquisition unmistakably vary results and could dictate actions taken in response. Neutrality is difficult to achieve, as humans are still heavily involved in the process. This is of value to the report because it reflects how biases are an inherent part of current algorithms and the artificial intelligence based these algorithms. This limits access to information and user autonomy.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Bryson, Caliskan, and Narayanan address the current prejudice evident in text semantics and its propagation into machine learning. They used a machine learning model and statistical analysis to demonstrate how word-embedding replicates sexism and racism seen in human culture. These biases are a result of using human language, and its connotations, as a base for machine learning. The implication of this accrued subjectivity is that human faults will be translated to and perpetuated by artificial intelligence. This compromises AI's objectivity and could further disadvantage those already affected by human discrimination. This article will be used to better demonstrate how bias is present in computers and artificial intelligence and the ethical ramifications. An example seen is the association of women with domestic roles in word-embedding and the potential ramifications.

Dadich, S. (2016). *The president in conversation with mitâĂŹs joi ito and wiredâĂŹs scott dadich.* https://www.wired.com/2016/10/president-obama-mit-joi-ito-interview/. (Accessed: 2017-11-19)

Danks, D., & London, A. (2017, 08). Algorithmic bias in autonomous systems. *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*, 4691-4697.

Danks and London examine the significance of bias in autonomous systems' algorithms and consider its weight in various situations. The achievement of objectivity in artificial intelligence will be determined by the mitigation of biases in its algorithms. Yet it is difficult to both define, identify and eliminate said biases. The unintentional emergence of racial prejudice in artificial intelligence demonstrates this fact. It is presented how bias can arise due to limited or erroneous input data, distorted methods of estimation, or inappropriate application. The removal of biases is very difficult due to the complexity of the algorithms, and human ethics. They caution labelling all biases as detrimental and address how bias may arise anywhere in algorithms, thanks to humans or not. The report will incorporate this article's material to emphasize how artificial intelligence, just as humans, is vulnerable to case-dependant bias.

Hirayama, N., Yoshino, K., Itoyama, K., Mori, S., & Okuno, H. G. (2015). Automatic speech recognition for mixed dialect utterances by mixing dialect language models. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *23*(2), 373–382.

In this article, Hirayama & al. propose a new automatic speech recognition system. The significance of this new system is its improved ability to recognize and interpret various dialects. This is crucial in oral human-computer interactions as computers often have difficulty understanding non-standardized language. They achieved such a system by renouncing earlier methods based on specific dialect lexicons to create a more pliable sort of system based on statistically derived translation rules to combine dialects. This information is relevant to the report as it reveals the unintentional bias present in computers and artificial intelligence's linguistic interactions with humans. It helps underline how individuals' access to artificial intelligence can be unjustly limited by their own circumstances, how this can hinder their advancement, and how it could be resolved.

Lecun, Y., & Cortes, C. (1999). *The MNIST database of handwritten digits.* (Retrieved from http://yann.lecun.com/exdb/mnist/)

Leenes, R., Palmerini, E., Koops, B.-J., Bertolini, A., Salvini, P., & Lucivero, F. (2017). Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues. *Law, Innovation and Technology*, *9*(1), 1–44.

This article discusses the problems with regulating AI use based on four major issues: creating laws that keep up with advancing technology; stimulating innovation without destroying of personal and social values; protecting and destroying social norms from robot bias; and, differentiating between a concern and an advancement to create proper limitations on AI advancements. We will use this source to discuss challenges with creating new guidelines for AI use in the fourth section of our report about future guidelines. We will specifically use this source to focus on creating laws and a code for limiting the functionality and usage of AI over a legitimate concern.

Liu, H.-Y. (2017). Irresponsibilities, inequalities and injustice for autonomous vehicles. *Ethics and Information Technology*, *19*(3), 193–207.

The legal and ethical framework of autonomous vehicles remains unclear, the author discusses the moral implications of the negative outcomes of autonomous vehicles. Liu suggests that the key lies in three factors which are: questioning the necessity for ascribing responsibility, differentiating between the disparate concepts that together inform existing notion of responsibility, approaching the issue from the polar perspectives of targeting and risk distribution. The utility of centralized risk distribution paradigm is explored, Liu discusses the benefits such as transforming the process to one of risk allocation rather than risk distribution. This is to allocate the risks and crash in a way that minimizes the costs. The autonomous vehicle will be used as an example to clarify the role of risk in negative outcomes.

National Science and Technology Council (U.S.) and Office of Science and Technology Policy and United States. (2016). *Preparing for the future of artificial intelligence* (Tech. Rep.). Executive Office of the President.

This federal report from the National Science and Technology Council's committee on AI outlines the current state of AI and its growing potential. The report describes course of action and recommendations that could be implemented by the federal government and other agencies to monitor and regulate the development of AI. This report also shows that some governments have already stated considering AI development as a major step in technology advancement that requires new laws to be implemented. We plan to use this report to give examples for recommendations in the future guidelines section of our report.

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

This paper discusses the imminence of powerful AI systems and their potential dangers. It mentions how the development of these technologies are being pushed by economic and military interests. It also describes how many AI systems can have similar 'meta-goals' that apply broadly to many goals. One example of this would be self-preservation-like behavior, which an AI may view as necessary in order to continue working towards its main goal. We plan on using this paper to discuss how hard it can be to ensure that AI will function as intended.

Scherer, M. U. (2016). Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, *29*(2), 353.

Scherer highlights the issue of a regulatory vacuum within artificial intelligence legislation and stresses the significance of tackling the issue. He proposes a regulatory Artificial Intelligence system: Artificial Intelligence Development Act (AIDA) which discusses how to best minimize the risks associated with autonomous Artificial Intelligence decision making. AIDA creates an agency that processes AI certifications that protect only certified sellers, manufacturers, and designers through a limited tort liability. While this regulatory regime is not yet a complete blueprint, its strong tort based system argues that it would push designers, programmers, and manufacturers to take responsibility for their Artificial Intelligence projects, developments and possible harms while also examining the safety system provided. This proposal can be used to examine the potentials of AI while ensuring its harms are kept to a minimum. Legislatures can thus be encouraged to develop more polished and detailed regulations to allow for the safe development and research of AI.

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*.

This report is a review of the One Hundred Year Study on Artificial Intelligence which began in 2014. This report reviews the potential advances and social challenges which arise from AI including areas of ethics, economics, and design over the last year. The study's goal is to assess AI growth and provide guidance for AI development including policies and systematic design to ensure AI is beneficial for the society. The report focuses on the future impact until the year 2030. We plan to use this report in the last section of our report to show a proposed impact in the future and recommendations for changes in public policy which is included in this study based on the future impact of AI on society.