Artificial Intelligence Ethics: Governance through Social Media

Jorge Buenfil[1], Ross Arnold[2], Benjamin Abruzzo[3], and Christopher Korpela[4]

Abstract— **A proposal is presented to facilitate machine self-learning of ethical behavior via human-curated training using online human behavioral data such as that found on social media and related sites. The proposed training data set is a mixture of human behavioral data found on social media and related sites that exhibit a wide variety of both ethical and unethical behavior which can help an artificially intelligent machine make ethical decisions during the process of solving real-world problems. The rapid proliferation of artificial intelligence (AI) applications worldwide highlights the need for normativity to protect individual rights, such as privacy, and the promotion of the common good; in other words, ethics. Governance of such widespread applications of AI as speech recognition, facial recognition, tracking of individuals using their personal electronic devices, etc., is needed to prevent abuses of such technologies by corporations or national governments. This paper presents a systemic view of the complexity of using principle-based governance to promote the ethical use of AI without unnecessarily hindering technological innovations needed to advance the state of the art in AI technology.**

I. INTRODUCTION

Homeland security and humanitarian assistance and disaster relief (HADR) operations demand that our tools and techniques be updated and improved constantly to better coordinate and execute the appropriate tasks. Artificial intelligence (AI) is one of the tools that has proven itself in many areas and holds great potential for use in homeland security and HADR operations.

Within the United States Department of Defense, AI refers to the ability of machines to perform tasks that normally require human intelligence for example, recognizing patterns, learning from experience, drawing conclusions, making predictions, or taking action - whether digitally or as the smart software behind autonomous physical systems[1]. Artificial intelligence (AI) is the capacity of machines to learn, reason, plan, perceive, and act; the primary traits we associate with human cognition (but, notably, not with consciousness or conscience). Although there are many types and degrees of AI and its boundaries are not always well-defined, AI systems generally exhibit behavior that resembles intelligence and/or the ability to make decisions. Many AI systems not only process data, but also learn from it and become smarter as they evolve over time.

AI systems that exhibit the highest degree of adaptability and sophistication are often applications of a style of computer programming called Probabilistic Programming or, most commonly, Deep Learning (DL). Deep learning methods are extensions of a mathematical structure known as a neural network. A neural network is a simplified representation of how a biological network of neurons would function. By mimicking the way neurons work, computers have made astounding advances in the fields of computer vision, natural language processing, speech recognition, and language translation. As a result, the application of neural networks and machine learning has given computer systems the ability to learn with minimal supervision, recognize complex patterns, and make recommendations and decisions on our behalf. Areas such as object recognition have seen rapid advancements since the development of deep learning, particularly after the arrival of Deep Convolutional Neural Networks (DCNN) [2]–[4]. These new abilities are being leveraged in many industries and sectors ranging from transportation to education to homeland security and national defense [5], [6].

Those advances are currently being applied to homeland defense and HADR. However, comparatively little attention has been placed on the implications to society stemming from the indiscriminate application of those technologies that may be required to make life and death decisions, such as which geographic areas to prioritize for emergency aid and personnel recovery. With great power comes great responsibility; despite the many advances in artificial intelligence achieved over the past several decades, the question remains: who or what is overseeing AI to behave in a manner that is acceptable to human beings as a whole?

We postulate that perhaps the answer to this question lies not in the wisdom of just one or even in a group of brilliant people, but in the whole of society. We suggest that AI systems leverage the collective knowledge and zeitgeist manifested in social media, provided this knowledge represents an acceptable, inclusive, representation of the population that will both enjoy the fruits of the application of ethical AI and also risk its negative consequences.

[1]J. Buenfil is a systems engineer at the U.S. Army CCDC-AC. jorge.r.buenfil.civ@mail.mil

[2]R. Arnold is a senior research engineer at the U.S. Army CCDC-AC. ross.d.arnold4.civ@mail.mil

[3]B. Abruzzo is a research engineer at the U.S. Army CCDC-AC. benjamin.a.abruzzo.civ@mail.mil

[4]C. Korpela is the Director of the Robotics Research Center, United States Military Academy. christopher.korpela@westpoint.edu

The views expressed are those of the authors and do not reflect the official policy or position of the US Army, Department of Defense, nor the US Government. This work was supported by the CCDC Armaments Center, the USMA Robotics Research Center, the Combating Terrorism Technical Support Office, and the Office of Naval Research.

Pamela McCorduck reflected the view of many technologists when she stated that she would rather take her chances with an impartial computer [7], implying that the results of an AI-developed decision-making process would be unbiased. Unfortunately, machine learning algorithms play into our existing biases and reaffirm them without our conscious realization [8]. It is quite possible that an AI application might discriminate against less-healthy or less-affluent people because its algorithms focus primarily on statistical averages or pattern recognition that favors the survival of the fittest [9], [10]. AI applications must be trained to check and correct biases in the data they use for analysis and learning. To the best of our knowledge the mechanisms to do that are not generally included in AI applications today. As algorithms guide more important facets of our lives, we need to trust, and deliberately design, machines that will treat us fairly and guide us toward the best possible version of humanity.

II. BACKGROUND

As Artificial Intelligence increases its penetration into everyday life, it becomes necessary to consider potential negative impacts of this relatively new technology. This must be done not just to protect the public in general, but also to protect the technology itself, which risks being restricted and or banned if the negative impacts of its use spiral out of control. An example of one such restriction is Europe's General Data Protection Regulation (GDPR), which took effect on May 25, 2018. The GDPR is the strictest approach to data protection yet devised. A very similar measure, the California Consumer Privacy Act, will take effect in California on January 1st, 2020. Another example is the ban on policy use of facial recognition enacted as a city ordinance in San Francisco, CA, in May 2019. The whole state of California is now considering a ban on police of use of any biometric surveillance system, including tattoos, gait and other individually distinguishable physical characteristics [11].

Interest in the topic of AI Ethics is increasing, which is evidenced by the previously mentioned legislative actions as well as the creation of several centers dedicated to the study of this topic. For example, the Center for Human-Compatible Artificial Intelligence (CHAI) is a multi-institution research group based at University of California, Berkeley. CHAI's goal is to develop the conceptual and technical wherewithal to reorient the general thrust of AI research towards provably beneficial systems[1]. Another center researching these issues is the Future of Life Institue through it's AI Safety Research program[2].

Additionally, some national governments are also producing the first guidelines to provide governance to AI systems. One such example is Singapore. In January 2019, the Personal Data Protection Commission (PDPC) of Singapore put forth a proposal towards these goals [12]. Their proposal centers on two guiding principles:

[1]https://humancompatible.ai/

[2]https://futureoflife.org/ai-safety-research/?cn-reloaded=1

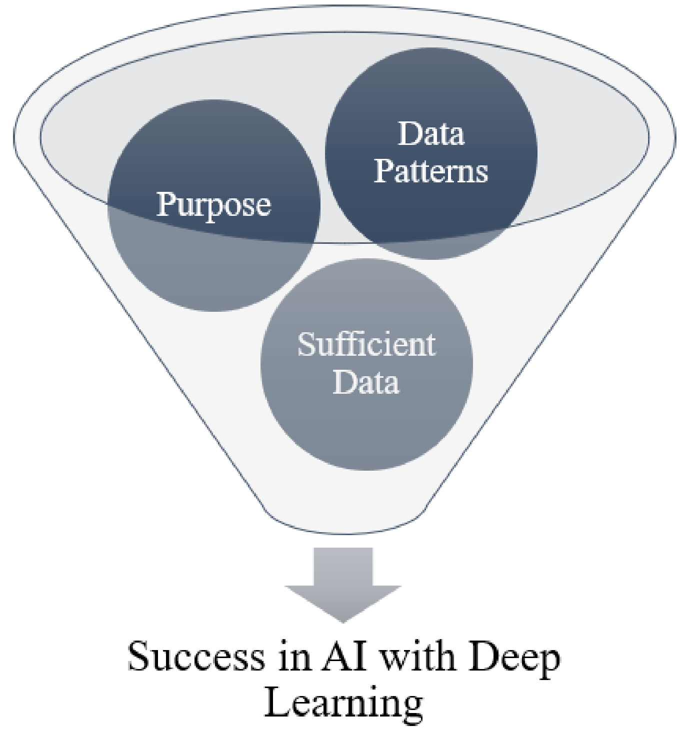

Fig. 1. Three ingredients for success in AI when using deep learning

1) Organizations using AI in decision-making should ensure that the decision-making process is explainable, transparent and fair.
2) AI solutions should be human-centric.

Even though some of the concepts are not necessarily well-defined, such as "fairness," the presence of these guidelines is representative of the concerns and motivations of governments to regulate AI. This framework is voluntary at present. Depending on the acceptance of these proposals, it is possible that Singapore may proceed to make their governance framework mandatory in the future.

A. Mechanisms of AI

Artificial Intelligence using deep learning requires three main ingredients to be successful (Fig. 1):

1) Data Patterns: Without these, predictions would be impossible;
2) Purpose: Absence of a mathematical formula to obtain the results of the search; otherwise machine learning has no purpose as the problem can be solved using the math formula
3) Sufficient Data: Sufficient quantity and type of data is needed to train the AI to recognize the patterns desired to normalize its behavior.

Data has been described as the fuel that drives AI. That is perhaps the main motivation behind corporations seeking to obtain as much data as possible via smartphones, desktop applications, online media, social engineering, and many other means. The data collection activity itself may not stem from malicious intent; however the users who might reasonably assume that they own their data do not necessarily know that. Oftentimes the data collected is not necessarily revealing when viewed in isolation, yet when aggregated with location information, date and time stamps, frequency of use, nearby people, etc., the data can infringe on privacy. This risk is analogous to the Operations Security (OPSEC) concept within the Defense Department that the collection of separate pieces of unclassified data can aggregate to a classified level.

As Dr. Kai-Fu Lee states: In deep learning, there's no data like more data [13]. The more examples of a given phenomenon a network is exposed to, the more accurately it can pick up patterns and identify traits of the real world. Given a significantly better set of data, an algorithm designed by a handful of intermediately skilled engineers usually outperforms one designed by a world-class deep-learning researcher.

Dr. Andrew Ng, Stanford University professor, and AI researcher at both Google and Baidu, uses the analogy that AI is the new electricity [14]. Dr. Ng believes that just as electricity transformed almost everything 100 years ago, AI will do the same in the next several years. Following that idea, Dr. Kai-Fu Lee has offered the analogy that if artificial intelligence is the new electricity, big data is the oil that powers the generators [13].

If these authors are to be believed, AI is fueling a revolution in the way we live our lives, work, consume news, travel, and more. According to Kai-Fu Lee, there are four waves to this AI revolution [13]:

1) Internet
2) Business
3) Perception
4) Autonomous

Internet AI processes data to understand user behavior and sell advertising and leverage additional sales by showing similar products to the ones people buy. The Google search engine is, arguably, one of the greatest AI systems that has yet been built.

AI helps businesses operate at a higher level of efficiency, organize their resources, manage inventories and resupplies, and many other useful tasks. High-frequency trading is another example of business AI. Algorithmic high-frequency traders account for more than half of the equity shares traded on US markets. Perception AI digitizes our physical world, learning to recognize faces, voices, understand requests, and perceive the world around us.

An interesting byproduct of the AI revolution is that AI naturally gravitates toward monopolies. Its reliance on data for improvement creates a self-perpetuating cycle: better products lead to more users, those users lead to more data, and that data leads to even better products, and thus more users and data. When monopolistic power is granted to AI companies, regulation of their activities to protect the common good become a necessity. Typically the free market is supposed to be self-correcting, but these self-correcting mechanisms break down in an economy driven by artificial intelligence.

The time to decide and implement AI governance, including ethical normativity, is now. The number of AI systems and robots is growing at an exponential rate. One estimate states that in 2010 the population of robots reached 8.6 million units. It is hard to estimate the total number of robots in 2019, but the growth has certainly been extensive. Available statistics account for mostly industrial robots in terms of robot density per 10,000 manufacturing employees. In 2017 at the top of the chart was the Republic of Korea with 710, followed by Singapore with 658, Germany 322, Japan 308, Sweden 240, Denmark 230, and the USA with 200.

In addition, the use of AI in the contexts of defense and homeland security opens the potential for a new type of risk [10]. When implemented on defense platforms, AI-driven autonomous systems can select and attack targets in ways that are faster and vastly more efficient than those performed by humans. Due to the potential for unintended collateral damage caused by these systems, the United States Department of Defense does not permit lethal fully autonomous weapons systems at this time. All weapons that include artificial intelligence must also include a human supervisor, or human-in-the-loop, for decision-making [15].

Any time a machine is empowered with the ability to make or influence decisions that affect peoples lives, ethics becomes an important factor in system development and deployment [16]. Even in humanitarian efforts, if an autonomous robot swarm is used to detect and report the locations of survivors [10] or used to identify different objects or people [6], [17], issues may exist that cause bias in reporting. Some research shows that computer algorithms developed by humans are easily contaminated by bias [16]. Certain people or types of people may be over-represented with regards to others. For example, automated face recognition techniques tend to be more effective on certain ethnic groups [18].

Perceptions of the ethical issues surrounding AI and autonomous systems also differ in different parts of the world. In Europe and North America, concerns about the use of autonomous Unmanned Aerial Vehicles (UAVs) tend to include invasion of privacy, misuse by government, and fears of an aviation accident [19]. However, concerns in the Kenya, where humanitarian drones were field tested, revolved around practical questions such as the strength of the UAVs camera, and how far the system could operate [20]. Though the drones in Kenya were not equipped with AI, the same concept applies. Given this knowledge, it is important to consider the concerns of the local communities with regard to AI governance, rather than to superimpose the concerns of certain nations in the mistaken assumption that the concerns are identical in other nations or communities.

These issues are not new, but they may be underestimated. To our knowledge, one of the most robust efforts to promote and develop friendly AI in such a way to benefit humanity as a whole, is OpenAI Inc. Founded in December 2015 with a pledge of $1 billion dollars by wealthy individuals from Silicon Valley such as Elon Musk [8] and an additional billion dollars invested by Microsoft in 2019 [21], OpenAI is

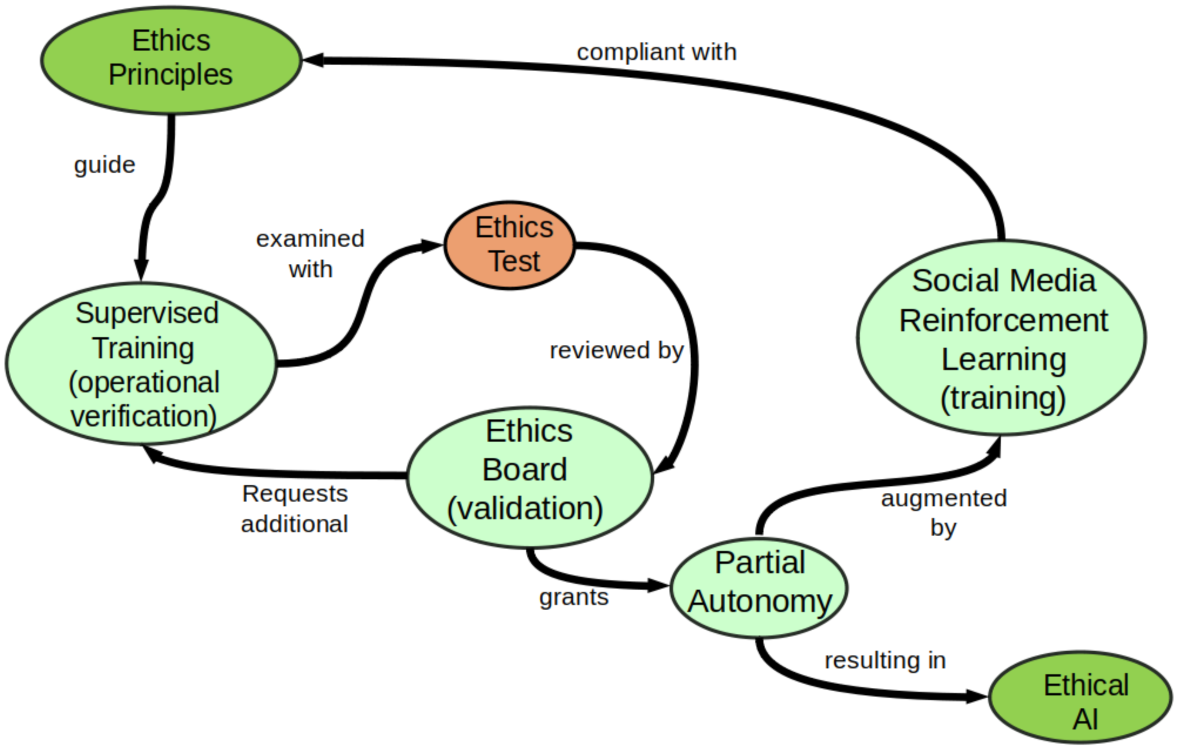

Fig. 2. AI Ethics Systemigram

a non-profit organization conducting research into artificial intelligence.

Given the issues stated and the difficulty in devising an overarching set of policies that govern all AI systems across all nations, we propose an approach in which the AI teach themselves ethics using public data sources. The results of the training would be overseen by regulatory mechanisms to prevent exploitation and flawed training.

III. PROPOSAL

We propose a method to approach AI governance by semi-autonomous machine learning through the use of human behavioral data such as social media, records of court cases, and related available data sources rather than explicit declarative regulatory controls and constraints. However, to prevent the ”Tay (bot) effect”, where the AI could be maliciously trained on social media the way the Microsoft chatter bot was [22], we introduce some regulatory mechanisms as can be seen in Fig. 2.

The main idea of this ongoing research is to apply a natural language processing algorithm that pulls insights out of data publicly available in textual format as is the case of reports on humanitarian actions, reports on military or criminal misconduct in a variety of clearly ethical situations with a socially acceptable outcome, as well as data on social media venues such as Twitter, Facebook, and Instagram. This algorithm then fits the results to ethics rules that could augment the core programming of the intelligent system.

A core philosophical consideration for this effort is the simple but quite challenging question of where to find data which would be generally considered ethical by an archetypal human population. In a broad sense, where and when does ethical behavior emerge in humans? If those situations, environments and contexts can be identified (and relevant data extracted), then such data may be used to train AI to act ethically in similar situations. However, even in the face of difficulty identifying these ideal situations, a committee of human ethicists could effectively curate and review decisions made by artificially intelligent systems in simulated scenarios and help to ensure ethical decisions as part of the AI training process.

There are three main phases for inserting ethics considerations into an AI system (Fig. 3).

1) Training
2) Operation
3) Reinforcement

In the Training phase we start by embedding into the system a set of basic ethics principles which cannot be altered. The principles to use may vary depending on the intended purpose of the system, but they should be very broad and self-evident to humans, such as those outlined by IEEE Ethically Aligned Design[3] or the 4 principles selected by the Statement of Ethical Principles for the Engineering Profession, proposed by the United Kingdom's Engineering Council and the Royal Academy of Engineering [4].

[3]https://ethicsinaction.ieee.org
[4]https://www.engc.org.uk/professional-ethics

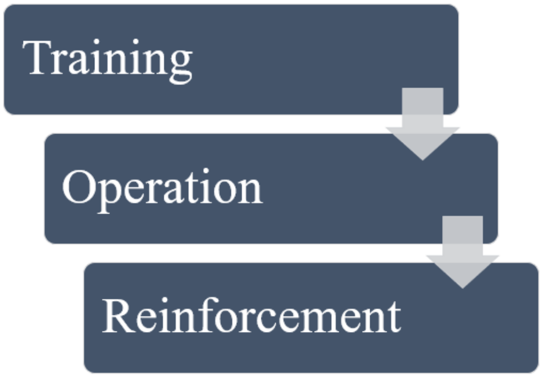

Fig. 3. Three phases of training are considered when inserting ethical behavior into AI systems.

Conceptually, an AI system would start with only those core inviolable ethical guidelines. Initial operation would require a human in the loop, who would provide supervision and feedback to teach the machine the correct ethical behavior in each selected scenarios. These scenarios could occur in a simulated environment as training begins, and slowly migrate to real scenarios as the training progresses. The machine would continuously add to its repertoire of ethical decisions based on feedback from the human trainer. After a certain amount of time, or when the machine has accumulated a predefined number of rules, it would be ready to submit to an Ethics Test. The results of the test would be evaluated by an Ethics Board for validation of the machine's ethical behavior.

If the results of the ethics test were unsatisfactory, the AI system would be sent back for further operational verification training with a human in the loop. The human would be consulted by the machine every time it reaches a low-confidence threshold. When the human supervisor determines that the AI system is ready to try to pass the ethics test again, it would be re-examined by the Ethics Board.

Once the results were satisfactory, the AI system would be allowed partial autonomy, with a human on-the-loop (rather than in-the-loop), acting as a supervisor but not interfering with the AI system unless it made an obvious mistake of judgement. At this point the AI system would be expected to demonstrate a basic level of Ethical AI. This level would be expected to increase over time. This approach reflects the advice provided by Dr. Arati Prabhakar, 20th Director of DARPA, that adaptive regulations allow one to experiment and learn without going too far [8]. The policies and regulations should achieve a degree of consensus and then provide stability so individuals and companies can count on a set of ground rules for a certain period of time.

IV. DISCUSSION AND OPEN RESEARCH CHALLENGES

A. Data Representation, Loss Functions, and Training

A key problem when training an AI system to perform ethically is how the ethics rules and guidelines are represented, especially when said rules are expected to expand by autonomous learning of new guidelines by making inferences on social media data. One way to do implement this is by using semantic web technology, based on formal ontologies to capture the essence of the concepts of interest that are likely to be found in social media channels. Ontologies are similar to taxonomies of ideas, where two concepts are interrelated by a link representing their relationship [23]. This data representation mechanism can be implemented via the Web Ontology Language (OWL) in a database such as Mongo, which is very extensible and capable of defining complex data structures [24].

B. Verification and Validation

We propose three steps for Verification and Validation (V&V):

1) Computer simulation, possibly with Agent-Based Modeling (ABM), to investigate how the system responds to stimuli from a simulated social media environment and how it captures ethical guidelines from those interactions.
2) Basic prototype with hardware in the loop, where we test a hardware architecture that could handle the separation of ethical principles into three logical locations:
    a) Core principle that are permanently burned into Read-Only Memory (ROM);
    b) Read-only files requiring root permissions to write. These files would contain the approved expanded guidelines that the system can use to make determinations of course of action to take;
    c) Files with read/write permissions that the system can use to stored ethical information derived from interactions with social media. These files are not to be used to influence courses of action, because they have not been vetted by the supervisory committee; until then they are only ethical guideline candidates.
3) Full autonomy in a controlled environment in which the AI system can use the hardware architecture described above to use derived ethical guidelines even without being approved by the ethical supervisory committee. The derived ethical guidelines can only be used when they do not contradict the core principles stored in ROM, nor the read-only directives that have already been approved by the supervisory committee. The only limitation to the AI's actions in this step would be the kinds of effects those actions can cause on the real world. These actions would be limited to presenting inferences via text, speech or graphics without access to actuators that physically impact the environment such as automatic doors, fire suppression systems, or anything that might risk physical harm or property damage.

C. Reinforcement Learning Beyond Social Media

Once the basic mechanisms of our proposal have been sufficiently tested and proven as trustworthy, it might be possible to allow the AI system to proceed by reading books, Wikipedia, specialized papers, conferences on Ethics, and other such sources by itself in order to survey a wider collection of cultural normativity. This would be effective especially if efficient mechanisms can be devised to allow the AI system to keep up to date on new publications in the field, including conference papers. More work needs to be done in this area.

V. CONCLUSION

Although many sources of AI governance are currently under investigation, we believe the application of social media data through our proposed AI self-learning process can serve a new, unfulfilled role in the governance of future artificial intelligence systems. Following our process, an AI system will be exposed to human behavioral data of various degrees on the ethical spectrum and receive feedback from a human trainer. Through several phases of learning, with an Ethics Test serving as a hard barrier between each, the presence of a human trainer in the loop will slowly be relaxed and the AI system allowed a higher degree of decision autonomy. Ultimately, the goal is a resultant Ethical AI system that can further its own self-learning in a useful way while adhering to strict, inviolable ethical guidelines established during the earlier phases of training. Further exploration is planned to bring this concept towards logical and physical architectures that can begin the all-important move towards the full realization of Ethical AI.

REFERENCES

[1] "Summary of the 2018 department of defense artificial intelligence strategy." [Online]. Available: https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.299.205

[3] C. Szegedy, A. Toshev, and D. Erhan, "Deep Neural Networks for Object Detection," in *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, pp. 2553–2561. [Online]. Available: https://papers.nips.cc/paper/5207-deep-neural-networks-for-object-detection.pdf

[4] L. Tobias, A. Ducournau, F. Rousseau, G. Mercier, and R. Fablet, "Convolutional Neural Networks for object recognition on mobile devices: A case study," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 12 2016, pp. 3530–3535. [Online]. Available: http://ieeexplore.ieee.org/document/7900181/

[5] B. Abruzzo, K. Carey, C. Lowrance, E. Sturzinger, R. Arnold, and C. Korpela, "Cascaded Neural Networks for Identification and Posture-Based Threat Assessment of Armed People," in *IEEE International Conference on Technologies for Homeland Security*, Woburn, Massachusetts, 2019.

[6] R. Arnold, B. Abruzzo, and C. Korpela, "Towards a Heterogeneous Swarm for Object Classification," in *IEEE National Aerospace and Electronics Conference*, Dayton, Ohio, 2019.

[7] P. McCorduck, *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. AK Peters/CRC Press, 2004.

[8] O. Groth and M. Nitzberg, *Solomons Code: Humanity in a World of Thinking Machines*. Pegasus, November 2018.

[9] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 77–91.

[10] R. D. Arnold, H. Yamaguchi, and T. Tanaka, "Search and rescue with autonomous flying robots through behavior-based cooperative intelligence," *Journal of International Humanitarian Action*, vol. 3, no. 1, p. 18, 12 2018.

[11] California Legislature, "AB-1215 Law enforcement: facial recognition and other biometric surveillance," 2019.

[12] "A proposed model ai governance framework," personal Data Protection Comission, Singapore. [Online]. Available: https://www.pdpc.gov.sg/Resources/Model-AI-Gov

[13] K.-F. Lee, *AI Superpowers: China, Silicon Valley, and the New World Order*. Houghton Mifflin Harcourt, 2018.

[14] S. Lynch, "Andrew Ng: Why AI is the New Electricity," *Insights by Stanford Business*, 2017.

[15] United States Department of Defense, "Autonomy in Weapons Systems," 2012.

[16] K. B. Sandvik, M. Gabrielsen Jumbert, J. Karlsrud, and M. Kaufmann, "Humanitarian technology: a critical research agenda," *International Review of the Red Cross*, vol. 96, no. 893, pp. 219–242, 3 2014. [Online]. Available: http://www.journals.cambridge.org/abstract_S1816383114000344

[17] R. Arnold, A. Kopeikin, B. Abruzzo, and C. Korpela, "Towards a General-Purpose, Replicable, Swarm-Capable Unmanned Aircraft System," in *IEEE International Conference on Technologies for Homeland Security*, Woburn, Massachusetts, 2019.

[18] N. Sharkey, "The impact of gender and race bias in AI," 8 2018. [Online]. Available: http://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/

[19] R. Arnold, K. Carey, B. Abruzzo, and C. Korpela, "What is a Robot Swarm: A Definition for Swarming Robotics," in *IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference*, New York, NY, 2019.

[20] D. Boyd, "Humanitarian Drones: Perceptions vs. Reality in Kenyas Tana Delta," The Sentinel Project, Tech. Rep., 2014. [Online]. Available: https://thesentinelproject.org/2014/11/18/humanitarian-drones-perceptions-vs-reality-in-kenyas-tana-delta/

[21] C. Metz, "With $1 billion from microsoft, an a.i. lab wants to mimic the brain," New York Times. [Online]. Available: https://www.nytimes.com/2019/07/22/technology/open-ai-microsoft.html

[22] H. Reese, "Why microsoft's 'tay' ai bot went wrong." [Online]. Available: https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/

[23] R. Arp, B. Smith, and A. Spear, *Building Ontologies with Basic Formal Ontology*. MIT Press, July 2015.

[24] H. Abbes and F. Gargouri, "M2onto: an approach and a tool to learn owl ontology from mongodb database," in *International Conference on Intelligent Systems Design and Applications*. Springer, 2016, pp. 612–621.