Spring 2018

Algorithms and Automation: Fostering Trustworthiness in Artificial Intelligence

Andrew B. Ware
University of New Hampshire, Durham, aw2009@wildcats.unh.edu

Algorithms and Automation: Fostering Trustworthiness in Artificial Intelligence

Keywords
artificial intelligence, technology, trustworthiness, transparency, explainability, AI

Subject Categories
Science and Technology Law | Science and Technology Studies | Technology and Innovation

Algorithms and Automation:
Fostering Trustworthiness in Artificial Intelligence

Andrew B. Ware

Advisor: Nicholas J. Smith

2018

I Introduction

The advancement of artificial intelligence (AI) presents humanity with opportunities as well as challenges. AI could contribute to increases in efficiency that radically impact productivity, and that may someday mitigate or even eliminate scarcity—offering abundance and ending wealth disparity. However, this future is not inevitable. AI is a powerful technology, and its power continues to grow. Like any powerful tool, it can be used to positive or negative ends. In developing and deploying AI systems, we must carefully consider the implications of this advancement and emphasize a collaboration between humans the technology.

In this paper I will argue that transparency and explainability are essential to maximizing the likelihood that AI has a positive impact on humanity. I articulate the importance of these features in the context of a collaboration between humans and AI, expressing that it is critical that we emphasize awareness and trustworthiness through meaningful transparency and explainability. Requiring that AI is able to adequately express to humans the process by which it makes determinations will foster reliance; this explanation is central to a productive partnership between AI and humans. Further, I propose that state regulation will be essential to steer AI toward beneficial ends and to monitor, prevent, and correct various abuses. Socialist principles appear best suited to establish and enforce features of transparency and explainability in emerging AI. Ultimately, it is important that we develop and deploy AI systems thoughtfully, carefully pursuing advancement to the extent that AI can be used as a tool that enhances human productivity and allows humanity to flourish.

I.I Background and Context

AI can be defined in many ways—for the purposes of this paper it refers primarily to machine learning (ML) algorithms that develop over time without being explicitly programmed.[1] Algorithms are a set of rules to be followed; in the case of ML, software algorithms are trained with data inputs and guidance from humans regarding desired outcomes. In this paper, use of the term "algorithms" refers to ML algorithms. These algorithms aim to optimize for a given metric—whether it be a more efficient use of agricultural resources like water on a farm that uses AI to monitor crop health and yield, a more accurate prediction of songs that might align with your musical taste when playing music with Spotify, or an effective determination of what content would be most engaging on Facebook to maximize clicks and advertising revenue.[2] It is challenging to evaluate these algorithms systematically in a way similar to how we evaluate algorithms that are not associated with ML, as outcomes are not similarly predetermined. In the case of algorithms that do not employ ML, a specific result is intended to be achieved, whereas ML algorithms are expected to produce an optimized result that is essentially unknown. As such, fostering trustworthiness through meaningful transparency and explainability will be particularly important in the development of AI.

Throughout this paper, "automation" refers to the implementation of algorithms and pre-programmed machines in an environment where they are used as a tool to increase productivity (the efficiency of production). These tools can be deployed to either augment

---

[1] SAS Institute, Inc, "Machine Learning: What is it and why it matters," *SAS,* 2018, https://www.sas.com/en_us/insights/analytics/machine-learning.html.

[2] Bernard Marr, "Spotify using Deep Learning to Create the Ultimate Personalised Playlist," *The Future Agency,* August 1, 2015, http://thefuturesagency.com/2015/08/01/spotify-using-deep-learning-to-create-the-ultimate-personalised-playlist/.

human labor (for example, in car manufacturing plants) or to offer guidance toward making better-informed decisions more quickly (through data analysis and insights derived from recognizing patterns). Even for instances in which automated machines replace specific tasks, these technologies have a more nuanced impact than simply the displacement of laborers. As automation is associated with greater productivity, the demand for labor increases for remaining complementary tasks.[3] In this paper, the term "machines" is sometimes used as substitute for "automated systems."

The scope of this paper includes weak or narrow AI—an implementation of AI technology that is focused on a specific task or set of tasks. The limits of this category of tasks are not easily determined; narrow AI is best understood in contrast with strong AI or artificial general intelligence (AGI), defined as "machine intelligence with the full range of human intelligence."[4] Further, practical implementations of AI today are best considered as weak.[5] While this distinction could become irrelevant after the advent of AGI, I note it to clarify my argument, specifying that I am discussing present technology and near-future advancements rather than what currently might be more accurately considered science fiction. Throughout this paper, I discuss AI as a tool that is informed by data and that is used by humans.

In discussing the significance of transparency and explainability, I will illustrate how productivity can increase and outcomes can improve as a result a collaboration between

---

[3] Daron Acemoglu and Pascual Restrepo, "Artificial Intelligence, Automation and Work," *MIT Economics,* January 4, 2018, https://economics.mit.edu/files/14641.

[4] Ray Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*, (United States: Viking Penguin, 2005), 260.

[5] Current examples of AI that I reference throughout this paper are therefore considered weak AI. AGI, while thought by many experts to be an inevitable future advancement, has not yet been achieved and is not immediately imminent.

humans and AI. Algorithms can process vastly more information than humans, in turn greatly enhancing the quality and velocity of our decision-making—benefiting farmers, doctors, corporate executives, government officials, and essentially everyone using human judgement to make determinations.

Further, automation can contribute to significant increases in efficiency and accuracy that are associated with higher levels of productivity. It seems evident that there will be benefits that result from the use of AI as a tool, as shown by current examples that range from improved manufacturing processes to enhanced medical diagnoses. However, it is not obvious that all AI applications produce a net benefit, nor that the benefits will be distributed equally. While AI may offer increases in efficiency and accuracy, what exactly algorithms optimize for—the objectives for which these powerful tools are used to achieve—is not inherent to the algorithms but in fact determined by the designers and developers of AI systems. Fostering trustworthiness through transparency and explainability will therefore be important in increasing the likelihood that these systems will be beneficial. However, it is important to note that this trustworthiness—achieved in part by meaningful transparency—will not necessarily ensure trust. In fact, there is the potential for trust to be undermined if AI systems are transparent about innerworkings, means, or ends that are malicious or with which people disagree. The distinction between trustworthiness and trust are significant, but are not the focus of this paper.

A military AI weapons targeting system, in which weapons are automated and algorithms are implemented to eradicate an enemy, is a salient example: either of two conflicting sides with competing interests and objectives could deploy these systems. Given the potential low cost and high power of such systems, small terrorist groups or even individual

bad actors might find AI to be an effective means of leveling the playing field in combat with more powerful enemies. In less extreme cases, the value-neutral nature of technology can still have a significant effect and remains an incredibly important consideration. Algorithms and automation alone are not ethically aligned, but are designed by humans to meet particular ends and reflect values represented in historical data. It is intrinsic to AI that it captures human and historical bias—despite sometimes seemingly more objective, algorithms are not independent from those creating, deploying, and using them.

I.II Bias in Data and Algorithms

The issue of algorithmic bias is a serious one—while AI presents promising potential in that it can consider vast volumes of information, operate incredibly efficiently, and recognize patterns in data that are not perceivable to humans, algorithms also incorporate implicit and explicit bias. As algorithms use data that are historic, outputs reflect past social and political injustices and inequalities. Consequently, AI systems might be unable to provide entirely unbiased or objective suggestions.[6] An example that illustrates how prejudice can perpetuated by algorithms is the use of criminal assessment tools.[7] Throughout the United States, software is used to predict the probability that someone convicted of a crime might reoffend—helping to make sentencing decisions in courtrooms by identifying which defendants are of a low enough risk that they could be released on bail while awaiting trial. This analysis is traditionally conducted with highly fallible judicial "instincts," so utilizing AI to process offender variables

---

[6] Jesse Dunietz, "The Fundamental Limits of Machine Learning," *Nautilus*, September 20, 2016, http://nautil.us/blog/the-fundamental-limits-of-machine-learning.

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

into an algorithm to predict recidivism seemed to be a promising means of improving the courts' predictive powers.

While the initial intention of the implementation of this software was to minimize overcrowding in jails, the algorithm correlated race with criminal history and incorrectly predicted that black defendants would reoffend nearly twice as often as it incorrectly indicated that white defendants would reoffend.[8] This software, called COMPAS, was not designed to predict higher recidivism rates for certain demographics—in fact, it was expected to provide a more objective assessment that was blind to factors such as race. By analyzing data on incarcerated individuals including past offenses, sentencing details, history of recidivism, and other factors, the system seeks to recognize patterns and determine the likelihood of recidivism of a given individual. Ideally, these factors would provide a demographic-independent prediction. However, in many instances, the guidance offered by the software is biased and inaccurate. For example, while white offenders sell and use drugs at a higher rate than non-whites, they face lower incarceration rates than non-whites as a result of systemic biases in law enforcement and the judicial system.[9] Since black defendants have faced higher incarceration rates in the United States, the algorithm that is being trained by this historical data is skewed—falsely indicating the rates at which individuals will reoffend based on race. While the company that develops the system denies such bias, it is evidently discriminatory nonetheless—while seemingly unintentional, prejudices present in the historical data were amplified throughout

---

[8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[9] Jeffrey Reiman and Paul Leighton, *The Rich Get Richer and the Poor Get Prison: Ideology, Class, and Criminal Justice*, (New York, New York: Routledge, 2017, Eleventh edition).

the process of training the algorithm. In the end, the case illuminates the danger in relying on AI systems to assist in decision making that results from bias in data and algorithms, and illustrates that transparency and explainability should be central to these systems. If there is an awareness of the issues associated with AI, there will be a clearer path forward in developing and deploying AI that is beneficial to humans.

In addition to this relatively subtler form of bias, algorithms can also be used to ends that might not be considered broadly beneficial. If developed by a private company that aims to maximize profits and shareholder value, for instance, an algorithm that offers agricultural guidance regarding which crops to grow and how to best grow them could be used to manipulate rather than to contribute to the achievement of the UN's Sustainable Development Goals, which are associated with advancing humanity and ensuring sustainability.[10] For example, an AI system in the context of agricultural resource management could use inputs including weather updates, geographical location, the nutrient content in soil, and other information in order to provide insight to farmers who currently rely upon a synthesis of historical recollection, empirical observations, tradition, and intuition. Better-informed farming could lead to higher crop yield, less water usage, and the more efficient use of other resources—and could even ensure food security. Despite this positive potential, though, developers could use data and algorithms to justify the exploitation of the land of a region. Algorithms could be

[10] Several of the UN's Sustainable Development Goals, which are set by the organization as objectives to achieve by 2030, could be addressed (at least in part) by advances in AI technology that is applied to the management and distribution of resources. For example, AI systems that offer resource usage guidance could present a contribute to ending world hunger and encouraging responsible consumption and production. [United Nations Development Programme, "Sustainable Development Goals," *United Nations*, January 2016, http://www.undp.org/content/undp/en/home/sustainable-development-goals/.]

designed to indicate an "optimal" distribution can be achieved by growing economically lucrative crops like sugarcane, providing guidance that most efficiently contributes to that end rather than offering and output associated with growing sustainable, nutritious foods. Evidently, algorithms can be aimed at the short term benefits of a small group in a way than seriously undermines the longer term benefits of humanity. In this sense, algorithms can seem to be an especially powerful form of capital that accelerate the disparity of wealth and well-being that emerges from profit-seeking. While there is the potential for AI systems to lead to more ideal outcomes, there is also the risk that systems will benefit some interests more than others, either intentionally or unintentionally.

To maximize the likelihood that the advancement of AI contributes to a sustainable future of decreased wealth disparity and increased well-being, there must be an awareness of bias and international regulation must be pursued that sets limits and guidelines regarding the development of AI, particularly by private, for-profit companies. In fostering this awareness, in establishing these regulations, and in advancing the technology itself, a broad range of perspectives must be considered—including not only the expertise of engineers, designers, and data scientists, but also the contributions of relevant non-governmental organizations, international groups, and government departments as well as the insight of academics, users of the systems, and policy- and law-makers. It should be communicated clearly to the public by international governments and organizations that algorithms cannot be separated from those developing and deploying AI systems, and that bias cannot be entirely eliminated from the process of human decision-making or the output of AI systems. To reconcile this reality, it is

critical that transparency and explainability are central to the advancement of AI and that trustworthiness is emphasized.

II Fostering Trustworthiness

AI systems must be trustworthy to ensure that the technology is used to contribute productively to humanity. Without trustworthiness, people will not be willing to engage with these systems. While humans are not especially trustworthy, it is challenging to accept guidance unless we believe it to be derived legitimately and credibly.[11] Essentially, people seek to understand, not completely but rather to an extent, how relevant conclusions are reached and how predictions are made that might impact their lives. The specific extent of explanation viewed as adequate varies among individuals, but humans do not generally rely on blind trust in decision-making and in evaluating situations. In some cases, it should be noted, we rely on determinations that are made by highly educated experts with extensive experience, engaging with these individuals without a very extensive understanding because we have no choice (e.g. in the determination of a medical diagnosis by a medical professional).

Suggestions offered by an AI system will likely not be considered very seriously if a system is not determined to be trustworthy. The foundational basis of the output of a system is rarely expressed explicitly in AI systems, which could seriously undermine trustworthiness. Neural networks—the foundation of many AI systems—function in a way that is implicit and opaque,[12] without clear reasons offered to explain outputs. They are often understood to operate

---

[11] Frank Alcock, David Cash, William C. Clark, Nancy M. Dickson, Noelle Eckley, and Jill Jäger, "Salience, Credibility, Legitimacy and Boundaries: Linking Research, Assessment and Decision Making," *KSG Working Papers Series RWP02-046*, February 3, 2003, https://dash.harvard.edu/handle/1/32067415.
[12] Will Knight, "The Dark Secret at the Heart of AI," *MIT Technology Review*, April 11, 2017, https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.

as a black box, or an entity that converts inputs to outputs without making its internal operations visible. As people would lack an awareness of the reasoning that underlies suggestions made by AI systems, the systems would not be considered trustworthy and people would likely hesitate or even refuse to accept the guidance offered by these systems. While explainability should be emphasized, other aspects of trustworthiness must be considered as well.

In considering trustworthiness, Onora O'Neill offers a useful framework. O'Neill suggests that trustworthiness can be understood by three primary features— competency, reliability, and honesty.[13] While it is not obvious that this understanding of trustworthiness applies in the context of the advancement of AI, I think it provides a reasonable way of approaching the issue of trustworthiness in AI systems.

II.I Evaluating Accuracy

These systems are shown to be competent by providing accurate results, or outcomes that are roughly aligned with expectations. However, it must be noted that since these systems are able to analyze a vastly greater amount of data than humans and are able to recognize patterns in this data beyond human perception, expected outcomes that are not necessarily the best benchmark for competency. While results may ultimately be accurate, it might not be realized that guidance is productive until far into the future. The complex multivariable analysis may initially appear faulty, since humans will likely be unable to foresee as far ahead as these algorithms. For example, in a game of chess against a computer, the system can analyze

[13] Onora O'Neill, "Trust, Trustworthiness, and Transparency," *EuroPhilantopics*, 2015, http://www.efc.be/human-rights-citizenship-democracy/trust-trustworthiness-transparency/.

each move by evaluating potential subsequent moves (through the end of the game), whereas even the best human chess players have a limit to their capacities to predict the myriad of future moves and related outcomes. A computer can predict the outcome eight or twelve or more moves in advance—therefore a recommendation offered by AI to move a certain piece might seem off to a human player who cannot comprehend the strategic permutations foreseen by the AI system. The recommended move might appear obviously wrong to a human even though it is far superior when viewed with the superior foresight of the AI system. In many applications, AI will likely be able to foresee outcomes months or years into the future—offering guidance based on these outcomes with justification that cannot be fully understood or realized by humans. When predictions are not able to be corroborated by humans, it will be challenging to determine accuracy.

Due to the aforementioned challenges, the competency of AI systems would best be measured in relation to the competency of humans: rather than pursuing an arbitrary objective standard of competency, these systems should be evaluated as being less, similarly, or more accurate than humans in making decisions, offering guidance, and providing predictions. For instance, if an algorithm is employed to assist in a medical diagnosis, it would be considered competent if it provided guidance that was ultimately correct (i.e. it accurately determines a diagnosis after analyzing the condition of a patient), at a rate at least as high as the success rate of humans performing similar tasks and making similar decisions. It should be noted, though, as in the example of a chess game introduced previously, that human expectations and actual accuracy will not necessarily be aligned. Since AI can consider so much more data than humans and can perceive patterns that humans cannot, it will not be possible in many instances to

evaluate accuracy in the context of human performance. This standard of competency will instead serve as guidance for the development of simpler AI systems. For more complex systems, the features of reliability and honesty will be more significant in fostering trustworthiness.

Since the accuracy of AI systems cannot be entirely assessed in direct relation to human performance, competency must be considered in the context of how humans currently make determinations—a notion explored later in this paper, in the section detailing transparency. In addition, because systems that rely on ML essentially evolve over time, their accuracy improves with use. Ultimately, it will be important that AI systems display competency by providing guidance and outputs that compare to human performance. While this comparison is not necessarily straightforward, seeking to understand accuracy in a human context and emphasizing features of reliability and honesty will contribute to fostering trustworthiness.

II.II Ensuring Consistency

Closely related to the evaluation of competency of these systems is another feature of trustworthiness previously outlined: reliability. Reliability is defined as the quality of performing consistently well.[14] Essentially, it describes the degree to which a system is dependably accurate or *competent over time*. Further, reliability is understood as the overall consistency of a statistical measure—to be characterized as a reliable result (or for a system to be recognized as reliable), similar results must be achieved under consistent conditions.[15] Computer systems are, by design, very consistent. Relative to the ability of humans to perform

[14] *Merriam-Webster*, *s.v.* "reliability," https://www.merriam-webster.com/dictionary/reliability.
[15] Ibid.

consistently, algorithms are often much more successful in producing unchanging results. The inputs that are analyzed by ML algorithms are controlled, whereas determining every factor that influences human behavior and decision-making is practically impossible.

While it may seem that algorithms and automated systems will therefore be expressly reliable, it should be noted that the nature of ML algorithms is such that they change over time with new data, even throughout their use. This makes it less clear that results would be completely consistent. It is challenging but important to ensure that ML contributes to a positive evolution rather than a degradation in the quality of results offered by AI systems. To resolve this issue, though, a broad range of perspectives—including those of individuals developing and deploying systems and those of individuals using and affected by these systems—must be involved in the advancement of AI beginning at the earliest stages of the process. For example, the perspectives of everyone from data scientists to climatologists to farmers should be incorporated throughout developing and deploying a system to manage agricultural resources and offer crop guidance.

Though synthesizing all of the viewpoints and the interests of a broad range of individuals will be challenging, a socialist approach would make it manageable. While profit incentivizes innovation in a capitalist system, this profit incentive would be minimized or even eliminated if the state oversees advancement. If the state mandates that a certain objective be achieved and dedicates resources to this end, it encourages socialist ideals of cooperation and collaboration among and between the public and private sectors. The competition of capitalism would be absent if this approach were taken, but the immense impact on humanity presented by AI would compensate for the lack of this motivation to innovate. It must be noted that this

impact could be positive or negative, leading to greater wealth and well-being or to the end of humanity as we know it: algorithms may manage agricultural resources much more efficiently than we do today and contribute to ending hunger, but automated weapons can also be used to decimate entire populations more efficiently than ever before. However, these drastically different ends both serve to catalyze advancement.

Instead of the potential for profit as the motivation to advance AI, the technology would be further developed and deployed to best align with the interests of a broad range of individuals, groups, and perspectives if regulated effectively. Collaboration and coordination among engineers, designers, data scientists, non-governmental organizations, international groups, government departments, academics, users of the systems, and policy- and law-makers—motivated by the incredible potential presented by the technology—will maximize the likelihood that AI systems will operate reliably. Also, to ensure systems are kept in check, mechanisms by which humans can directly intervene with AI systems to make modifications must be in place, and people must be able to continuously monitor systems.

In addition to influencing the operation of systems with new data, people must be able to alter systems to accurately reflect and respect the ever-changing social and economic structures, values, and practices of the cultural contexts in which these systems are used. Eventually, it is likely that AI will be used to moderate systems and to evaluate reliability. In the earlier stages of development, emphasizing a collaboration between humans and AI will be significant in ensuring that systems are used safely and serve to benefit humanity.[16] Ultimately,

[16] Daron Acemoglu and Pascual Restrepo, "Artificial Intelligence, Automation and Work," *MIT Economics,* January 4, 2018, https://economics.mit.edu/files/14641.

though, the remaining feature of trustworthiness—honesty—remains important in contributing to the measurement of competency, the evaluation of reliability, and the use and engagement of systems.

II.III Meaningful Transparency

Even if competency is shown by accuracy and reliability is established by consistent performance, these systems must demonstrate honesty. In the context of AI, honesty is best interpreted at transparency. This interpretation is reasonable because meaningful transparency would ensure that everyone who engages with and who is affected by AI has a practical understanding of how systems make determinations. Honesty is cultivated through open communication and explanation; by offering justification regarding behavior and judgements aligned with what people want to know and what it is important that they know, people can build honesty and subsequently become more trustworthy. AI systems would best demonstrate this honesty through transparency and explainability—in the context of algorithms and automation, these features can be understood to be practically equivalent to honesty.

As discussed earlier, the neural networks that form the foundation of many AI systems are inherently opaque and do not provide clear explanation for guidance that is offered. For people to accept this guidance, these systems must be developed and deployed in a way that emphasizes explainability and meaningful transparency—fostering an awareness of the reasoning that underlies suggestions. Honestly, or meaningful transparency, requires that

systems not merely express internal operations and calculations, but that explanations are widely accessible, comprehensible, and useful to users, serving to enhance understanding.[17]

Several projects are being pursued to design AI systems to be honest, embedding explainability and meaningful transparency in these systems throughout development and deployment. For example, the United States Defense Advanced Research Projects Agency (DARPA) has established an initiative to create Explainable AI (XAI), reflecting the importance of machine's abilities to explain decisions and actions to humans. Ultimately, DARPA states that this is essential if people are "to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners."[18] The project illustrates that meaningful transparency—achieved by explainability—is central to trustworthiness. While establishing the initiative and beginning research to this end is a significant step forward, it is important to note that details regarding the level of transparency and extent of explanation have not been clarified. Again, it is important to go beyond just overcoming opaqueness and to provide useful information about the underlying processes of algorithms and automated machines to a broad range of users.

For even the most seemingly simple tasks completed by algorithms, such as that of sorting email and distinguishing junk or spam messages, certain levels of transparency and explanation would not be adequate for all users. While exposing the data points that are used and how they are employed to determine whether or not a message is legitimate may be a

---

[17] Simon Beard, "Will AI Help to Build a Fairer World? The Answer Is in Our Hands," *The Huffington Post (UK)*, November 21, 2017, www.huffingtonpost.co.uk/entry/will-ai-help-to-build-a-fairer-world-the-answer-is-in-our-hands_uk_5a12f556e4b023121e0e950d.

[18] David Gunning, "Explainable Artificial Intelligence (XAI)," *DARPA*, August 10, 2016, https://www.darpa.mil/program/explainable-artificial-intelligence.

useful explanation to an expert or even someone familiar with the design of similar systems, it would likely not be possible for a general email user to understand. For users without the knowledge and expertise required to understand the design of AI systems, simpler explanations must be provided that allow for greater awareness and that contribute to trustworthiness. If just experts have access to an understanding of how systems operate, people would instead need to find these individuals to be trustworthy rather than finding the systems trustworthy. This is problematic in that it is indirect and undemocratic; people will lack confidence in the guidance offered by systems if interactions with AI are mediated and governed by a small group of limited perspectives, such as a few engineers, corporate executives, or government officials. To reconcile this issue, reasonable expectations should be set by regulatory bodies to ensure that systems can be understood by a wide range of people with a wide variety of backgrounds. In an effort to ensure safety and reduce risks associated with developing and deploying these systems, this high standard of regulation is necessary. Policy-makers, researchers, and engineers should collaborate and seek to understand the implications of the advancement of AI on those developing it, those engaging with it, and those affected by it.[19] This understanding will inform the level of explanation that is required, and can help to build the trustworthiness of AI systems.

A socialist approach will increase the likelihood that this regulation be established and enforced effectively. In capitalism, competition is emphasized and therefore anti-competitive regulation can be challenging to implement. While this can be beneficial in reducing prices of

[19] Miles Brundage, Shahar Avin, Jack Clark, et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *The Malicious Use of Artificial Intelligence*, February 21, 2018, https://maliciousaireport.com/.

products to consumers and increasing consumer choice, the risks associated with the advancement of AI are too serious to encourage competition rather than collaboration and coordination among a diversity of perspectives. Further, given that much of AI is proprietary intellectual property—an aspect of capitalism that allows for a competitive advantage to be maintained (at least temporarily)—it is particularly important that regulation be pursued. If certain organizations maintain control of the most powerful algorithms and earn profits under capitalism as a result, it will be challenging to achieve meaningful transparency. A democratic socialist approach that encourages organizations to communicate, collaborate, and to coordinate efforts through regulation will ensure trustworthiness and maximize the likelihood that is AI is pursued for the benefit of humanity.

In addition to the issue of determining an adequate degree of explanation for honesty considering various perspectives and the objective of fostering trustworthiness, it is also important to consider how explainability might be conceived if and when AI systems become more complex than any human is able to understand. While the advancement of AI should be pursued carefully and regulation should be implemented to ensure systems are productive, the limits of human understanding should not limit further technological advancement—science and technology can be used to positively or negatively impact humanity and potential risks should not deter us from achieving immense potential benefits. Rather than requiring every AI system that is developed and deployed to be transparent in explaining its innerworkings, other AI systems could be employed to explain more complex systems, and these systems that are designed to explain can subsequently provide meaningful transparency. Essentially, these systems would operate as interpreters between humans and advanced AI. Researchers are

beginning to explore this potential by developing algorithms to explain how systems make decisions.[20] If their approach is integrated into AI systems, algorithms would require systems to provide evidence-based justification and would offering natural-language explanations to users. These algorithms would overlay the algorithms that are to be explained, and would serve to shine light into the black box that characterizes neural networks. Ultimately, these algorithms would provide users an understanding of how outputs are derived from inputs within the AI system, and could be designed to explain how it determines this understanding as well.

While these efforts contribute to increased explainability, transparency remains a complicated objective that presents several challenges. In addition, there is no objective measurement that indicates the degree to which a system is transparent (or an obvious threshold at which a system might be considered adequately transparent).[21] These issues could be resolved by policy-makers though regulation; however, there are further obstacles to overcome. Depending on who is advancing AI—private companies, the government, or academic research groups, for instance—intellectual property is another consideration. There could be incentive not to disclose how algorithms operate, as doing so might undermine profit potential within capitalist economic structures, or even global power. It is unlikely that the systems will be developed in a way that is open-source (at least without regulation), such that source code is made available to the public and such that systems are developed in a way that is

---

[20] Zeynep Akata, Trevor Darrell, Lisa Anne Hendricks, Dong Huk Park, Marcus Rohrbach, and Bernt Schiele, "Attentive Explanations: Justifying Decisions and Pointing to the Evidence," *ARXIV*, December 2016, https://arxiv.org/pdf/1612.04757v1.pdf.

[21] Adrian Weller, "Challenges for Transparency," *ARXIV*, July 29, 2017, https://arxiv.org/pdf/1708.01870.pdf.

collaborative. These types of systems are already being developed and deployed by a broad range of organizations, and it is clear that ensuring that users understand how algorithms work is not a priority.[22] Here, the urgency of regulation cannot be understated. Guidelines regarding the advancement of AI, established by consulting a broad range of perspectives, will be critical in ensuring that explainability is emphasized and trustworthiness is fostered. Again, democratic socialist principles must govern the development of AI to ensure that regulations can be effectively established and enforced.

Despite these seemingly significant challenges, though, transparency and explainability should be put in a human context so that more reasonable expectations can be set. Just as there are issues with objectively measuring competency and reliability, it is more practical to evaluate the explainability of AI systems in relation to the honesty of people. Humans are unable to fully explain decisions or the justification for guidance—illustrated well by returning to the earlier example of medical diagnoses. In determining a diagnosis, a systematic method is used but does not entirely explain the process and information used in making the determination. A particularly complex diagnosis—incorporating a synthesis of evidence of the current patient, education, past observation and experience, and other experts' perspectives—might not be able to be explained to other medical professionals, let alone patients who require a simpler, more accessible explanation.

---

[22] In some instances, it seems that there is a priority to exploit data and to prevent users from understanding how algorithms operate. [Matthew Rosenberg, Nicholas Confessore, and Carole Cadwalladr, "How Trump Consultants Exploited the Facebook Data of Millions," *The New York Times*, March 17, 2018, https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html.]

As such, it would be reasonable to expect systems and to set regulations for AI to be at least as transparent and explainable as humans, even if complete awareness or communication is lacking. If AI can explain itself at least as well as humans explain their own actions and reasoning, systems would demonstrate transparency and honesty. Further, this explanation and meaningful transparency would contribute to reliability and allow for a reasonable evaluation of competency. If the expectations for explanation are set for machines similarly to our expectations of humans, trustworthiness could be fostered effectively. By assessing AI in a human context, at least until trustworthiness is built, humans could reasonably rely upon these systems just as they rely on each other—constantly ensuring that trust is maintained, but ultimately contributing positively to productivity.

Trustworthiness is critical to engagement with and utilization of these systems; explainability is central to fostering trustworthiness. This would ensure honesty and contribute to the evaluation of competency and reliability. What people want to know and what it is important that they know must be determined and expressed effectively—in the end, a thoughtful consideration of transparency will offer humans the opportunity to improve outcomes and increase productivity.

III A Productive Partnership

While there is the potential for AI to have a positive impact on humanity, issues remain once trustworthiness has been fostered. Even once systems demonstrate competency, reliability, and honesty through transparency and explainability, AI algorithms previously thought not to be dangerous can evolve to have malicious intent, can be hacked, and can violate the trust that has been built. Even with an awareness of biased objectives, it would be challenging to form a

consensus in many cases regarding the extent to which these prejudices should be overlooked if the output of systems proves to be more reliable than humans. If guidance offered is less biased than the guidance typically offered by people making decisions and offering guidance, should it be followed? Similarly, if initially insignificant issues are amplified over time as algorithms develop, when should we question whether the AI is still trustworthy? It should be noted that these issues are not drastically different from current issues we face with people and the tools upon which we rely. Regulation could resolve or at least mitigate some of these concerns, but ensuring that AI is trustworthy—while critical—is not sufficient.

In pursuing the advancement of AI, it is important that a collaboration between humans and AI that enhances human ability is emphasized. Rather than developing and deploying AI systems to replace humans in the completion of tasks, algorithms and automation should instead serve as tools that increase efficiency and augment human productivity.[23] While some have been vocal about concerns regarding the development of AI, sharing a cautious outlook, others involved in its advancement have indicated intense optimism that the long term benefits of AI will far outweigh its potential costs. For example, Microsoft CEO Satya Nadella has written that humans and AI can collaborate to "solve society's greatest challenges like beating disease, ignorance, and poverty."[24] It is important to consider and outline priorities moving forward that will ensure human and machine will work together for the benefit of humanity,

---

[23] Daron Acemoglu and Pascual Restrepo, "Artificial Intelligence, Automation and Work," *MIT Economics,* January 4, 2018, https://economics.mit.edu/files/14641.

[24] Satya Nadella, "The Partnership of the Future," *Slate*, June 28, 2016, www.slate.com/articles/technology/future_tense/2016/06/microsoft_ceo_satya_nadella_humans_and_a_i_can_work_together_to_solve_society.html.

and there is evidence that a collaboration between humans and AI will greatly improve outcomes and ultimately offer the opportunity to improve conditions for everyone in the world.[25] It should be reiterated, though, that this technology could be used by undemocratic states, small terrorist groups, or even individual bad actors as means to undermine democracy or efficiently target and attack large populations in violent conflicts. To minimize this risk and ensure that AI is developed and deployed safely, collaboration among a myriad of perspectives should be emphasized, in addition to cross-disciplinary research efforts that seek to understand potential mis-uses of AI.[26]

Efforts to ensure that AI collaborates effectively with humans are exemplified by the Partnership on Artificial Intelligence, a consortium of companies including Amazon, Apple, DeepMind, Google, Facebook, IBM, and Microsoft. The group was established in 2016 to "study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society."[27] Several of the thematic pillars around which the Partnership on AI is organized involve the interaction between humans and the technology: in addition to emphasizing "fair, transparent, and accountable AI," the organization has pillars of "collaborations between people and AI systems" and "AI, labor, and the economy."[28] The

---

[25] Future of Life Institute, "Asilomar AI Principles." *Future of Life Institute*, 2017, https://futureoflife.org/ai-principles/.
[26] Miles Brundage, Shahar Avin, Jack Clark, et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *The Malicious Use of Artificial Intelligence*, February 21, 2018, https://maliciousaireport.com/.
[27] Partnership on AI, https://www.partnershiponai.org/.
[28] Partnership on AI, https://www.partnershiponai.org/thematic-pillars/.

Partnership on AI illustrates that for AI "to benefit people and society" there must be focus on fostering trustworthiness though explainabilty as well as ensuring that a productive partnership between humans and AI is formed.[29] Further, there is evidence that supports the notion that collaboration—associated with augmentation and enhancement—is more productive than the replacement of human labor, guidance, or decision-making.

One example that shows the success of the collaboration between human expertise and AI is the identification of metastatic breast cancer.[30] A team consisting of members from Harvard Medical School and Massachusetts Institute of Technology developed ML algorithms to analyze images of lymph node biopsies and found that while the system was somewhat successful in identifying cancer, it did not exceed the ability or rate of accuracy of the diagnoses of human pathologists. However, when the output of the AI system was utilized by pathologists in the process of determining a diagnosis, the identification success rate was higher than that of the human diagnosis—human error was reduced by approximately 85 percent. While human ability and intelligence seemed to exceed that of AI, greater results were achieved when AI was employed as a tool by pathologists. In the end, the study established that through a partnership between humans and AI, "significant improvements in the accuracy of pathological diagnoses" could be achieved.[31] It is likely that similarly improved outcomes will become possible as AI systems are developed and deployed for other use cases.

---

[29] Partnership on AI, https://www.partnershiponai.org/introduction/.
[30] Andrew Beck, Rishab Gargeya, Humayun Irshad, Aditya Khosla, and Dayong Wang, "Deep Learning for Identifying Metastatic Breast Cancer." *ARXIV,* June 18, 2016, https://arxiv.org/pdf/1606.05718.pdf.
[31] Ibid.

Another example that indicates that AI might lead to mutually beneficial outcomes in the management and distribution of certain resources. The tragedy of the commons is a long-standing issue in economics—it is a theory that describes how, in a shared-resource system, individuals behave in a self-interested way that contrasts with the common good.[32] Since people do not often behave with community interests in mind, it is challenging to model efficient allocation of common pool resources. A DeepMind project offers a new perspective, though, that transcends existing economic and political solutions. Instead of relying on non-cooperative game theory, which often leads to the failure of agents to most efficiently allocate resources, DeepMind used ML algorithms to show that trial-and-error learning in common-pool resource appropriation can lead to socially positive cooperative outcomes.[33] It was found that when exclusion policies are easier to implement and when predictions are based on historical data representing individuals' interests and behaviors, appropriation decisions are less collectively beneficial and efficient. The conclusion: useful insight can be gained by incorporating AI into behavioral economic theories that aim to model human behavior and interaction. Rather than relying entirely upon AI to replace humans in making decisions, offering guidance, and providing predictions, fostering trustworthiness and working *with* AI will lead to more increased productivity.

---

[32] Partha Dasgupta and Veerabhadran Ramanathan, "Pursuit of the common good," *Science*, September 19, 2014, http://science.sciencemag.org/content/345/6203/1457.

[33] Charles Beattie, Joel Z. Leibo, Julien Perolat, Karl Tuyls, and Vinicius Zambaldi, "A multi-agent reinforcement learning model of common-pool resource appropriation," *ARXIV*, July 2017, https://arxiv.org/pdf/1707.06600.pdf.

In addition to these scenarios, algorithms and automation—when used as tools by humans—can increase the accuracy and efficiency of other tasks, particularly those that are repetitive.[34] Without being hindered by the biological needs and limitations of humans, AI systems can operate uninterrupted and can often operate more rapidly (and with greater accuracy as well as consistency). Humans would maintain an important role in this operation, though; people will need to monitor AI systems to ensure that outputs continue to reflect and respect cultural context and ideals, providing results with which we are satisfied. Further, if automation significantly increases productivity, the demand for labor would increase for complementary tasks that are not as easily automated.[35] Ultimately, a thoughtful collaboration between humans and AI, in which algorithms and automated systems are used as tools, along with regulation that ensures development is beneficial rather than detrimental to humanity, will result in improved outcomes associated with better decision-making and more efficient productivity.

IV The Future of Labor and Life

A greater reliance on algorithms and automation through the collaboration of humans with AI systems raises several questions regarding the future of labor and life that warrant brief discussion. As productive partnerships form between humans and AI systems across many industries, significant changes in work and the workforce will occur. Increases in productivity

---

[34] Mark Muro and Scott Andes, "Robots Seem to Be Improving Productivity, Not Costing Jobs," *Harvard Business Review*, June 16, 2015, https://hbr.org/2015/06/robots-seem-to-be-improving-productivity-not-costing-jobs.

[35] Daron Acemoglu and Pascual Restrepo, "Artificial Intelligence, Automation and Work," *MIT Economics,* January 4, 2018, https://economics.mit.edu/files/14641.

and output will follow from improved accuracy, decision-making, and efficiency. It will likely be necessary to redefine labor, which would have effects that extend into all aspects of life.

Through a collaboration with AI, there is the potential for efficiency to be increased to a degree that allows humanity to achieve abundance and virtually eliminate scarcity—mitigating people's need to work as many hours per week and reducing the reliance on labor to exchange for goods and services that fulfill needs and wants. John Maynard Keynes explored this technological utopia of radically increased efficiency nearly a century ago.[36] Keynes described a future in which the laborers who "do not sell themselves for the means of life…will be able to enjoy the abundance" and will enjoy "the art of life itself."[37] While this future may be distant or even not entirely achievable, it is reasonable to expect that AI will contribute to increased productivity as well as increased unemployment.

Unemployment and underemployment are generally negatively perceived (especially in the United States). However, if the rate of productivity and volume of output is able to be increased to a point where people's needs are able to be fulfilled without reliance on labor, this perception is likely to change. The capitalist structure—the economic and societal framework of the United States and many other nations (particularly those facing the most intense American influences and the highest levels of globalization)—requires a culture in which work greatly impacts identity and in which people are "reduced to selling their labor power in order to live."[38] As Marx articulated, a system that relies on wage labor for productivity results in people

---

[36] John Maynard Keynes, "Economic Possibilities of Our Grandchildren," *Essays in Persuasion*, 1930, https://www.marxists.org/reference/subject/economics/keynes/1930/our-grandchildren.htm.
[37] Ibid.
[38] Karl Marx and Friedrich Engels, *The Manifesto of the Communist Party*, Marxists Internet Archive – Marx Engels Archive, 1848, https://www.marxists.org/archive/marx/works/1848/communist-manifesto/.

losing their autonomy and becoming beholden to the capitalists who controls the means of production. Under capitalism, since labor is essentially exchanged for goods and services with currency, work is intrinsic to one's identity, and is even closely associated with what many people believe to be their purpose for living.

In the transition toward this future of less labor and more abundance, it is likely that a universal basic income (UBI) or some sort of universal compensation will be necessary. It would serve to bridge the gap between our current state of production and consumption under capitalism and a social and economic structure in which scarcity is not nearly as significant a consideration. It has been shown that people remain productive after UBI is implemented; however, I will not detail this transition and the many arguments it entails here. [39] Throughout the shift brought about by an increased reliance and collaboration with AI systems, people will inevitably need to change what they value and how they perceive themselves—socioeconomic distinctions may increasingly fade away. A future in which humanity's wealth and well-being is maximized is most closely aligned with socialism, in which the state regulates the infrastructure of algorithms and automation. If the advancement of AI is not approached in a way that is managed by the state and that incorporates a variety of viewpoints, it is likely that a small elite will maintain control of the technological infrastructure, drastically increasing wealth and well-being disparity and taking the exploitation of capitalism to its extreme end.[40]

---

[39] Basic Income Earth Network, http://basicincome.org/basic-income/.
[40] Peter Frase, *Four Futures*, (Brooklyn, New York: Verso, 2016).

Prior to Keynes's predictions and descriptions of a future in which people are no longer required to "sell themselves for the means of life,"[41] Marx articulated how he imagined a similar future would appear.[42] When reflecting on life under capitalism, Marx expressed that "activity is not voluntarily, but naturally, divided, man's own deed becomes an alien power opposed to him, which enslaves him instead of being controlled by him," and continued to explain that "as the distribution of labor comes into being, each man has a particular, exclusive sphere of activity, which is forced upon him and from which he cannot escape."[43] In illustrating the specialized nature of roles in capitalism that are often not entirely the result of individual choice, Marx was depicting a contrast to a future in which an individual can become "accomplished in any branch he wishes…to hunt in the morning, fish in the afternoon, rear cattle in the evening, criticize after dinner, just as I have a mind, without ever becoming hunter, fisherman, herdsman or critic."[44] While there is the potential for this future to encourage the pursuit of passions that contribute to productivity, there is the libertarian challenge to the socialist view. Rather than collaboration, cooperation, and coordination, the current libertarian capitalist perspective emphasizes competition and conflict as efficient instigators of innovation. According to the libertarian view, free markets can improve productivity and provide a

---

[41] John Maynard Keynes, "Economic Possibilities of Our Grandchildren," *Essays in Persuasion*, 1930, https://www.marxists.org/reference/subject/economics/keynes/1930/our-grandchildren.htm.
[42] While Marx was writing about communism—a system in which the means of production are controlled by the proletariat—rather than a technological future of increased efficiency, he ultimately envisioned a society and economy in which scarcity was virtually eliminated ("from each according to his ability, to each according to his needs") and in which labor and production was no longer directly related to life and consumption. [Karl Marx, *Critique of the Gotha Programme*, Marxists Internet Archive – Marx Engels Archive, 1875, https://www.marxists.org/archive/marx/works/1875/gotha/index.htm.]
[43] Karl Marx, *The German Ideology*, Marxists Internet Archive – Marx Engels Archive, 1845, https://www.marxists.org/archive/marx/works/1845/german-ideology/.
[44] Ibid.

motivation for innovation though profit incentive. Further, a capitalist structure is thought to encourage personal choice and preserve liberty and maintain equality while not establishing a homogeneous society.

There exists the potential for the advancement of AI to lead to a stagnant society in which there is limited incentive to innovate; however, like our views of work and its association with identity, this perception results from the current capitalist state-of-mind. As efficiency is increased with technology, people will begin to understand the value of coordination and collaboration in achieving further advancements. Reliance on AI systems, resulting from the trustworthiness of AI, will create an atmosphere in which collaboration is emphasized over competition. In describing a future of increased autonomy in labor and life, Marx suggests that what is understood to be the challenge of work-life balance in capitalism would become a work-life harmony. People would be free to pursue passions while remaining productive, and the concept of livelihood would be greatly altered. This, too, would also significantly contribute to a new conception of identity and purpose. Ultimately, if humans and AI collaborate, improvements in accuracy, decision-making, and efficiency will lead to increased productivity and could result in a future of abundance. These shifts in work and resource availability would impact all aspects of life, and serve to remove its close connection with labor.

V Conclusion

It is incredibly important that transparency and explainability are central to the advancement of AI, as these features will foster trustworthiness and a reliance on systems. Further, meaningful transparency will ensure an awareness of the innerworkings of systems similar to the justification provided by people regarding decisions, actions, and determinations.

Regulation by the state will be essential in maximizing the likelihood that AI will benefit humanity; a socialist system, emphasizing collaboration and coordination, will be most effective in establishing and enforcing regulations. A productive partnership between humans and AI should also be emphasized; AI can present a positive future if it is pursued thoughtfully.

Since the advancement of AI presents humanity with both opportunities and challenges, we must be cognizant of risks while pursuing development that contributes to a positive future of increased productivity. Through carefully considering the implications of the advancement of algorithms and automation as well as implementing regulations that require a broad range of perspectives be involved in developing and deploying systems, it will be more likely that AI is used in a way that is benefits humanity. Regulation will clarify the levels of explainability that should be expected of systems, and will outline how to effectively evaluate competency and reliability. Ultimately, the advancement of AI should be pursued transparently, and systems should be developed to feature transparency—while an ideal future is not inevitable, we can move toward greater productivity, abundance, equality, and well-being if AI is trustworthy and works in collaboration with humans.

Bibliography

Acemoglu, Daron and Pascual Restrepo, "Artificial Intelligence, Automation and Work," *MIT Economics*, January 4, 2018, https://economics.mit.edu/files/14641.

Akata, Zeynep, Trevor Darrell, Lisa Anne Hendricks, Dong Huk Park, Marcus Rohrbach, and Bernt Schiele, "Attentive Explanations: Justifying Decisions and Pointing to the Evidence." *ARXIV*, December 2016. https://arxiv.org/pdf/1612.04757v1.pdf.

Alcock, Frank, David Cash, William C. Clark, Nancy M. Dickson, Noelle Eckley, and Jill Jäger. "Salience, Credibility, Legitimacy and Boundaries: Linking Research, Assessment and Decision Making." *KSG Working Papers Series RWP02-046*, February 3, 2003. https://dash.harvard.edu/handle/1/32067415.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica*, 23 May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Armstrong, Stuart. "AI safety: three human problems and one AI issue." *Intelligent Agent Foundations Forum*, May 19, 2017. https://agentfoundations.org/item?id=1388.

Basic Income Earth Network. http://basicincome.org/basic-income/.

Beard, Simon. "Will AI Help to Build a Fairer World? The Answer Is In Our Hands." *The Huffington Post (UK)*, November 21, 2017. www.huffingtonpost.co.uk/entry/will-ai-help-to-build-a-fairer-world-the-answer-is-in-our-hands_uk_5a12f556e4b023121e0e950d.

Beattie, Charles, Joel Z. Leibo, Julien Perolat, Karl Tuyls, and Vinicius Zambaldi. "A multi-agent reinforcement learning model of common-pool resource appropriation." *ARXIV*, July 2017. https://arxiv.org/pdf/1707.06600.pdf.

Beck, Andrew, Rishab Gargeya, Humayun Irshad, Aditya Khosla, and Dayong Wang. "Deep Learning for Identifying Metastatic Breast Cancer." *ARXIV*, June 18, 2016. https://arxiv.org/pdf/1606.05718.pdf.

Beddoes, Zanny Minton. "Special report: The world economy - As You Were." *The Economist*, October 13, 2012. http://www.economist.com/node/21564413.

Bostrom, Nick and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." *The Cambridge Handbook of Artificial Intelligence*, 2014. https://intelligence.org/files/EthicsofAI.pdf.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford, England, United Kingdom: Oxford University Press, 2014.

Brokaw, Alex. "This startup uses machine learning and satellite imagery to predict crop yields." *The Verge*, August 4, 2016. https://www.theverge.com/2016/8/4/12369494/descartes-artificial-intelligence-crop-predictions-usda/.

Brundage, Miles, Shahar Avin, Jack Clarke, and others. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." *The Malicious Use of Artificial Intelligence*, February 21, 2018. https://maliciousaireport.com/.

Burke, Marshall, W. Matthew Davis, Stefano Ermon, Neal Jean, David B. Lobell, Michael Xie. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Stanford University Sustainability and Artificial Intelligence Lab*, 2016. http://sustain.stanford.edu/predicting-poverty/.

Cave, Stephen. "Artificial Intelligence: a five-point plan to stop the Terminators taking over." *The Telegraph*, September 30, 2016. http://www.telegraph.co.uk/technology/2016/09/30/artificial-intelligence-a-five-point-plan-to-stop-the-terminator/.

Centre for the Study of Existential Risk. http://cser.ac.uk/.

Dasgupta, Partha and Veerabhadran Ramanathan, "Pursuit of the common good." *Science,* September 19, 2014. http://science.sciencemag.org/content/345/6203/1457.

Davis, Nicholas. "What is the fourth industrial revolution?" *World Economic Forum*, January 19, 2016. https://www.weforum.org/agenda/2016/01/what-is-the-fourth-industrial-revolution/.

Doshi-Velez, Finale and Mason Kortz. "Accountability of AI Under the Law: The Role of Explanation." *Berkman Klein Center for Internet & Society at Harvard University,* November 27, 2017. https://cyber.harvard.edu/publications/2017/11/AIExplanation.

Dunietz, Jesse. "The Fundamental Limits of Machine Learning." *Nautilus*, 20 September 2016. http://nautil.us/blog/the-fundamental-limits-of-machine-learning.

Fairness, Accountability, and Transparency in Machine Learning. http://www.fatml.org/.

Felton, Ed and Terah Lyons. "The Administration's Report on the Future of Artificial Intelligence." *The White House,* October 12, 2016. https://www.whitehouse.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence.

Frase, Peter. *Four Futures*. Brooklyn, New York: Verso, 2016.

Friedenberg, Michael. "A.I. Ethics Emerge." *Cio* 28, no. 11 (September 2015): 6. *Business Source Complete*, EBSCOhost.

Future of Humanity Institute. https://www.fhi.ox.ac.uk/.

Future of Life Institute. "Asilomar AI Principles." *Future of Life Institute*, 2017. https://futureoflife.org/ai-principles/.

Google - PAIR - People+AI Research Initiative. https://ai.google/pair/.

Gunning, David. "Explainable Artificial Intelligence (XAI)." *DARPA*, August 10, 2016. https://www.darpa.mil/program/explainable-artificial-intelligence.

Keiper, Adam, and Ari N. Schulman. "The Problem with 'Friendly' Artificial Intelligence." *The New Atlantis*, no. 32 (2011): 80-89. http://www.jstor.org.libproxy.unh.edu/stable/43152658.

Keynes, John Maynard. "Economic Possibilities of Our Grandchildren." *Essays in Persuasion*, 1930. https://www.marxists.org/reference/subject/economics/keynes/1930/our-grandchildren.htm.

Knight, Will. "The Dark Secret at the Heart of AI." *MIT Technology Review*, April 11, 2017. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.

Kuang, Cliff. "Can A.I. Be Taught to Explain Itself?" *The New York Times,* November 21, 2017. https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.

Kurzweil, Ray. *The Singularity Is Near: When Humans Transcend Biology*. United States: Viking Penguin, 2005.

Leverhulme Centre for the Future of Intelligence. http://lcfi.ac.uk/.

Li, Fei Fei. "How to Make A.I. That's Good for People." *The New York Times,* March 7, 2018. https://www.nytimes.com/2018/03/07/opinion/artificial-intelligence-human.html.

Marr, Bernard. "Spotify using Deep Learning to Create the Ultimate Personalised Playlist." *The Future Agency*, August 1, 2015. http://thefuturesagency.com/2015/08/01/spotify-using-deep-learning-to-create-the-ultimate-personalised-playlist/.

Marx, Karl and Friedrich Engels. *The Manifesto of the Communist Party*. Marxists Internet Archive – Marx Engels Archive, 1848. https://www.marxists.org/archive/marx/works/1848/communist-manifesto/.

Marx, Karl. *A Contribution to the Critique of Political Economy*. Marxists Internet Archive – Marx Engels Archive, 1859. https://www.marxists.org/archive/marx/works/1859/critique-pol-economy/index.htm.

—. *Critique of the Gotha Programme*, Marxists Internet Archive – Marx Engels Archive, 1875. https://www.marxists.org/archive/marx/works/1875/gotha/index.htm.

—. *The German Ideology*. Marxists Internet Archive – Marx Engels Archive, 1845. https://www.marxists.org/archive/marx/works/1845/german-ideology/.

Matz, Cade. "Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free." *Wired,* April 27, 2016. https://www.wired.com/2016/04/openai-elon-musk-sam-altman-plan-to-set-artificial-intelligence-free/.

Muro, Mark and Scott Andes. "Robots Seem to Be Improving Productivity, Not Costing Jobs." *Harvard Business Review*, June 16, 2015. https://hbr.org/2015/06/robots-seem-to-be-improving-productivity-not-costing-jobs.

Nadella, Satya. "The Partnership of the Future." *Slate,* June 28, 2016. http://www.slate.com/articles/technology/future_tense/2016/06/microsoft_ceo_satya_nadella_humans_and_a_i_can_work_together_to_solve_society.html.

O'Neill, Onora. "Trust, Trustworthiness, and Transparency." *EuroPhilantopics*, 2015. http://www.efc.be/human-rights-citizenship-democracy/trust-trustworthiness-transparency/.

Pande, Vijay. "Artificial Intelligence's 'Black Box' Is Nothing to Fear." *The New York Times,* January 25, 2018. https://www.nytimes.com/2018/01/25/opinion/artificial-intelligence-black-box.html.

Partnership on AI. https://www.partnershiponai.org/.

Prescott, Bonnie. "Better Together: Artificial intelligence approach improves accuracy in breast cancer diagnosis." *Harvard Medical School*, June 22, 2016. https://hms.harvard.edu/news/better-together/.

Price, Huw. "Now it's time to prepare for the Machinocene." *Aeon*, October 17, 2016. https://aeon.co/ideas/now-it-s-time-to-prepare-for-the-machinocene/.

Reiman, Jeffrey and Paul Leighton. *The Rich Get Richer and the Poor Get Prison: Ideology, Class, and Criminal Justice*. New York, New York: Routledge, 2017 (Eleventh edition).

Rosenberg, Matthew, Nicholas Confessore, and Carole Cadwalladr. "How Trump Consultants Exploited the Facebook Data of Millions." *The New York Times*, March 17, 2018. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html

Russell, Stuart, Danial Dewey, and Max Tegmark. "Research Priorities for Robust and Beneficial Artificial Intelligence." *Association for the Advancement of Artificial Intelligence*, Winter 2015. http://futureoflife.org/data/documents/research_priorities.pdf.

Russell, Stuart, Sabine Hauert, and Russ Altman. "Robotics: Ethics of artificial intelligence." *Nature* 521, no. 7553 (May 28, 2015): 415-418. *Academic Search Complete*, EBSCOhost.

SAS Institute, Inc. "Machine Learning: What is it and why it matters." *SAS,* 2018. https://www.sas.com/en_us/insights/analytics/machine-learning.html.

Stocker, Michael. "Decision-making: Be wary of 'ethical' artificial intelligence." *Nature* 540, 525 (December 22, 2016). http://www.nature.com/nature/journal/v540/n7634/full/540525b.html.

"The world's most valuable resource is no longer oil, but data." *The Economist*, May 6, 2017. https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource.

United Nations Development Programme. "Sustainable Development Goals." *United Nations*, January 2016. http://www.undp.org/content/undp/en/home/sustainable-development-goals/.

United Nations Division for Sustainable Development. "Transforming our world: the 2030 Agenda for Sustainable Development." *United Nations*, October 21, 2015. https://sustainabledevelopment.un.org/post2015/transformingourworld/.

Vincent, James. "Google uses DeepMind AI to cut data center energy bills." *The Verge*, July 21, 2016. https://www.theverge.com/2016/7/21/12246258/google-deepmind-ai-data-center-cooling/.

Weller, Adrian. "Challenges for Transparency." *ARXIV,* July 29, 2017. https://arxiv.org/pdf/1708.01870.pdf.

Zakrzewski, Cat. "Musk, Hawking Warn of Artificial Intelligence Weapons." *The Wall Street Journal,* July 27, 2015. http://blogs.wsj.com/digits/2015/07/27/musk-hawking-warn-of-artificial-intelligence-weapons/.