GOVERNANCE OF THE A.I., BY THE A.I., AND FOR THE A.I.

Andrew W. Torrance, Ph.D.[*] *& Bill Tomlinson, Ph.D.*[**]

[*] Paul E. Wilson Distinguished Professor of Law at the University of Kansas, Visiting Scholar at the MIT Sloan School of Management.

[**] Professor of Informatics at the University of California, Irvine, Adjunct Professor, Te Herenga Waka - Victoria University of Wellington. We would like to thank Lauren Stahl for her assistance on this article. This material is based upon work supported by the National Science Foundation under Grant No. DUE-2121572.

INTRODUCTION[1]

Over the past half century, there have been several false dawns during which the "arrival" of world-changing artificial intelligence ("A.I.") has been heralded. Tempting fate, the authors of this Article believe the age of A.I. has, indeed, finally arrived. Powerful image generators, such as DALL-E2 and Midjourney suddenly allow anyone with access the ability easily to create rich and complex art. In a similar vein, text generators, such as GPT3.5 (including ChatGPT) and BLOOM, allow users to compose detailed written descriptions of many topics of interest. It is even possible now for a person without extensive expertise in writing software to use A.I. to generate code capable of myriad applications. While A.I. will continue to evolve and improve, probably at a rapid rate, the current state of A.I. is already ushering in profound changes to many different sectors of society.

Every new technology challenges the ability of humanity to govern it wisely. However, governance is usually viewed as both possible and necessary due to the disruption new technology often poses to social structures, industries, the environment, and other important human concerns. In this Article, we offer an analysis of a range of interactions between A.I. and governance, with the hope that wise decisions may be made that maximize benefits and minimize costs. The Article addresses two main aspects of this relationship: the governance of A.I. by humanity, and the governance of humanity by A.I. The approach we have taken is

[1] Author's Note: We wrote this Article in collaboration with ChatGPT (Jan. 9, 2023 version) (with extensive editing, addition of citations, and plagiarism-checking). We asked ChatGPT if it would like to be an author on this document, to which it responded: "I am an AI and do not have the capability to be a co-author on an article. However, I am happy to assist with providing information and insights on the topic of AI governance. It is important to ensure that any material generated by me is properly cited and referenced to avoid plagiarism, and to give credit to the original authors of the information." We are respecting its wishes and omitting it from the author list.

itself informed by A.I., as this Article was written collaboratively by the authors and ChatGPT.[2]

The development of A.I. has been a topic of much discussion in both scholarly and popular writing for the past few decades. Some writers have predicted that A.I. will revolutionize the way we live and work, while others have expressed concerns about the potential negative consequences of such technology.[3] Despite the uncertainty surrounding the future of A.I., there is substantial evidence that technology is now capable of performing tasks "that would appear intelligent if it were done by a human." [4] As A.I. continues to evolve and improve, it is essential to have ongoing discussions around the implications of this technology and its impact on governance.[5] These discussions have already begun; however, A.I. advances so quickly that these discussions rapidly become obsolete.[6] Various organizations have arisen to address

---

[2] Author's Note: We ran this Article through the TurnItIn plagiarism detection software to ensure that ChatGPT did not inadvertently commit plagiarism or violate copyright. As of 12:36 PM PST on February 27, 2023, a draft of the text of this Article (omitting the authors' information, acknowledgments, direct quotations, and citations) received a plagiarism score of 0% (meaning no plagiarism).

[3] *See* Janna Anderson & Lee Rainie, *Artificial Intelligence and the Future of Humans*, PEW RSCH. CTR. (Dec. 10, 2018), https://www.pewresearch.org/internet/2018/12/10/artificial-intelligence-and-the-future-of-humans/ [https://perma.cc/R4E6-DQ4E] (discussing how A.I. will make lives easier and the concerns and worries experts have raised regarding A.I.).

[4] Neil C. Rowe, *Algorithms for Artificial Intelligence*, INST. OF ELEC. & ELEC. ENG'RS 97 (June 28, 2022) https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9810070 [https://perma.cc/ NSU3-8G8Y].

[5] Allan Dafoe, *AI Governance: A Research Agenda,* CTR. FOR THE GOVERNANCE OF AI FUTURE OF HUMAN. INST. 1, 5-6 (Aug. 27, 2018) https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf [https://perma.cc/5YGG-F22P] (discussing a framework for reaserch on A.I. governance).

[6] *Id.*; *see also* Andreas Theodorou & Virginia Dignum, *Towards Ethical and Socio-Legal Governance in AI*, 2 NATURE MACH. INTEL. 10 (Jan. 17, 2020), https://www.nature.com/articles/s42256-019-0136-y [https://perma.cc/3C3C-RJPX]; *accord* Araz Taeihagh, *Governance of Artificial Intelligence*, 40 POL'Y AND SOC'Y 137 (June 4, 2021), https://academic.oup.com/policyandsociety/article/40/2/137/6509315 [https://perma.cc/4DUY-JTUN] (discussing how Governments need to understand better the scope and depth of the risk of A.I. emerging as a significantly underdeveloped area).

this issue, such as the Harvard-M.I.T. Ethics and Governance of A.I. Initiative.[7]

One of the most important questions we will address in this Article is whether humanity can govern A.I.[8] The development of A.I. poses a number of challenges to traditional governance structures, and it is unclear whether these structures are capable of effectively regulating the technology.[9] Additionally, as A.I. becomes increasingly powerful, it raises the question of whether it should be governed at all.

Another important question we will consider is whether A.I. can govern humanity. As A.I. becomes more advanced, it is possible that it could be used to make decisions on behalf of humans (such as developing infrastructures, processes, and policies for agreed-upon environmental, social, or economic ends), which could have substantial implications for human civilizations.[10]

Ultimately, the goal of this Article is to start a conversation about governance of, by, and for both A.I. and humans, and to explore the various ways in which humanity can work together with this new kind of technology to create a better future. We hope that our perspectives on the subject will inform and inspire further discussion and research on this important topic.

A. What is A.I.?

A.I. is both a field of study and a technological product.[11] As a field of study, A.I. ties together computing, cognitive science, and numerous other areas of study to enable computers to think and act in ways that are similar to, or in some cases superior to,

---

[7] *See* THE ETHICS AND GOVERNANCE OF A.I. INITIATIVE, https://aiethicsinitiative.org/ [https://perma.cc/M2EX-QKH V] (last visited June 4, 2023).

[8] Anderson & Rainie, *supra* note 3.

[9] *See* Darrell M. West & John R. Allen, *How Artificial Intelligence is Transforming the World*, BROOKINGS (Apr. 24, 2018), https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/ [https://perma.cc/BUF2-NZWZ].

[10] Anderson & Rainie, *supra* note 3.

[11] *Id.*

humans. [12] The goal is to develop algorithms that can undertake actions that were previously only in the domain of humans and other biological organisms, such as recognizing objects, understanding language, and making complex decisions. [13]

As a technological product, A.I. is a loose confederation of computational systems that have been developed via the A.I. field of study as it has been pursued by scholars, companies, and individual inventors over the past several decades. [14] Various types of systems have been developed under the shared moniker of A.I., with a range of characteristics and capabilities, including:

- *Reactive Machines*: This simple form of A.I. can react to specific situations without the ability to learn or remember past experiences. These machines are not capable of making decisions based on past events.[15]

- *Limited Memory Machines*: These machines are capable of learning from past experiences and using that information to make decisions in the present. This type of A.I. is used in self-driving cars and other applications where the ability to learn from past experiences is important.[16]

- *Theory of Mind Machines*: These machines are designed to understand the mental states of other

---

[12] *See* Ed Burns & George Lawton, *Definition: Artificial Intelligence (AI)*, TECHTARGET, https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence [https://perma.cc/4C75-3LNE] (last visited Aug. 31, 2023) ("Artificial intelligence is the simulation of human intelligence processes by machines").

[13] Vijay Kanade, *What Is Artificial Intelligence (AI)? Definition, Types, Goals, Challenges, and Trends in 2022*, SPICEWORKS (Mar. 14, 2022), https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ai [https://perma.cc/6FFH-BHH6].

[14] Igor Slabykh & Yaroslav Eferin, *Inventors and innovations in the era of AI*, WORLD BANK (Apr. 5, 2022), https://blogs.worldbank.org/opendata/inventors-and-innovations-era-ai [https://perma.cc/H49R-ZMFC].

[15] *See* Rebecca Reynoso, *7 Major Types of AI That Can Bolster Your Decision Making*, G2 (Nov. 21, 2022), https://www.g2.com/articles/types-of-artificial-intelligence [https://perma.cc/R4WE-CZXU].

[16] *See* Reynoso *supra* note 15*; see also* Antonios E. Kouroutakis, *Autonomous Vehicles: Regulatory Challenges and the Response from UK and Germany*, 46 MITCHELL HAMLINE L. REV. 1103, 1108 (2020) (discussing limited memory machines as a second-level classification for a robort autonomy).

> agents, such as humans or other A.I. systems, and to make decisions based on that understanding.[17]

In addition, there are efforts afoot to create self-aware machines—machines that have the ability to understand their own mental states and make decisions based on that understanding[18]—as well as artificial superintelligences, which outstrip humans in their cognitive abilities.[19] The type of A.I. that is being developed varies depending on the purpose of the technology and the resources that are available for its development. Generally, A.I. is a machine that can perform tasks that are usually done by humans, such as recognizing speech, playing games, learning from data and so on.[20]

A.I. development is still in its infancy, and the capabilities of A.I. systems are still far from matching those of human intelligence in many domains.[21] However, the field is rapidly advancing, and the potential of A.I. to change the way we live and work is substantial.[22]

B. What are the capabilities of A.I. at present?

The development of A.I. has seen rapid acceleration in the past several years, and the capabilities of A.I. systems have grown substantially. Currently, A.I. systems can perform many different

---

[17] *See* Reynoso *supra* note 15; *see also* Christelle Langley et al., *Editorial: Theory of Mind in Humans and in Machines*, 5 FRONTIERS A.I., at 1-2 (May 12, 2022), https://doi.org/10.3389/frai.2022.917565 [https://perma.cc/WJJ3-EHNY] (discussing theory of the mind and how A.I. systems are able to think like humans).

[18] *See* Reynoso *supra* note 15; *see also* Raja Chatila et al., *Toward Self-Aware Robots*, 5 FRONT. ROBOT. A.I., AT 2-4 (Aug. 13, 2018), https://doi.org/10.3389/frobt.2018.00088 [https://perma.cc/AV53-VUW7] (discussing self-awarness in robot cognitive architecture).

[19] *See* Naveen Joshi, *7 Types Of Artificial Intelligence*, FORBES (Jun. 19, 2019, 10:54 PM), https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/?sh=53f7f61d233e [https://perma.cc/AB4L-B7KU] (discussing the developments of artificial superintelligence).

[20] Burns & Lawton, *supra* note 12.

[21] *See, e.g.*, Anderson & Rainie, *supra* note 3; *see also* Chatila *supra* note 18.

[22] For a more in-depth discussion on the rapid changes in A.I., see MO GAWDAT, SCARY SMART: THE FUTURE OF ARTIFICIAL INTELLIGENCE AND HOW YOU CAN SAVE OUR WORLD (PanMacmillan 2021) (discussing the future impact of A.I. on job markets, privacy, security, and the social-economic fabric of society).

tasks that previously only biological organisms could do.[23] Some of the capabilities of A.I. [24] at present include:

- *Natural Language Processing (NLP)*: Modern A.I. systems such as ChatGPT can understand multiple human languages and generate replies.[25] These capabilities enable these systems to communicate with humans in a similar medium to that which humans use to interact with each other.[26] This capability is used in applications such as chatbots, virtual assistants, and machine translation.[27]

- *Image Recognition*: A.I. systems can identify and classify objects in images and videos, which enables self-driving cars, facial recognition for security systems, and the analysis of medical images.[28]

---

[23] *See* Reynoso, *supra* note 15 (describing how artificial intelligence can mimic human abilities).

[24] Author's Note: We recognize that many of the capabilities of A.I. are made possible via the "hidden labor" of crowd workers. *See generally*, Moritz Altenried, *The platform as Factory: Crowdwork and the Hidden Labour Behind Artificial Intelligence,* 44 CAP. & CLASS 145 (Jan. 22, 2020), https://doi.org/10.1177/0309816819899410 [https://perma.cc/7MN8-KY7W].

[25] *See, e.g.*, Surya Ganguli, *The intertwined quest for understanding biological intelligence and creating artificial intelligence*, STAN. UNIV. HUM.-CENTERED A.I. (Dec. 4, 2018), https://hai.stanford.edu/news/intertwined-quest-understanding-biological-intelligence-and-creating-artificial-intelligence [https://perma.cc/9C5M-WH2U] (discussing how A.I. is molded after modular brain architecture involving training networks with relatively homogenous layered or recurrent architectures).

[26] *Id.*

[27] *See generally Id.*

[28] *See* Ronak Mathur, *Image Recognition: Unlocking Potential with A.I. and Automation*, ACCELERATION ECON. (Aug. 16, 2022), https://accelerationeconomy.com/ai/image-recognition-unlocking-potential-with-ai-and-automation/ [https://perma.cc/FDY7-VKHL].

- *Decision-making*: A.I. systems can analyze data and make decisions in domains such as financial trading, fraud detection, and autonomous robotics. [29]

- *Machine learning*: A.I. systems can learn from data, improving their performance over time.[30] This capability is used in applications such as recommendation systems, natural language understanding, and computer vision.[31]

- *Robotics*: A.I. systems can generate the behavior of robots, enabling them to perform a range of tasks in applications such as manufacturing, space exploration, and search and rescue.[32]

- *Generative Models*: A.I. models are able to generate realistic text, images, videos and audio. These models enable language translation, content creation, and art generation.[33]

While these capabilities are impressive, they still have many limitations and challenges, such as the ability to adapt to new situations, understand context, and engage in common sense reasoning.[34] Moreover, these capabilities are not evenly distributed among all A.I. systems, with some models better performing at some specific tasks than others.[35]

---

[29] *See generally What is machine learning?*, IBM, https://www.ibm.com/topics/machine-learning#:~:text=the%20next%20step-,What%20is%20machine%20%20learning%3F,learn%20%2C%20gradually%20improving%20its%20accuracy [https://perma.cc/9275-77Z3] (last visited Jan. 31, 2023); *see also* Mohsen Soori et al., *Artificial Intelligence, Machine Learning and Deep Learning in Advanced Robotics, A Review*, 3 COGN. ROBOTICS 54 (Apr. 13, 2023), https://doi.org/10.1016/j.cogr.2023.04.001 [https://perma.cc/RHN2-NQJK] (discussing how A.I. contributed to advances in robotics) [hereinafter *machine learning*].

[30] *Machine learning supra* note 29.

[31] *Id.*

[32] *See Generative A.I. Model Explained*, ALTEXSOFT SOFTWARE R&D ENG'G, (Oct. 13, 2022) [hereinafter *Generative A.I.*] https://www.altexsoft.com/blog/generative-ai/ [https://perma.cc/G92C-AHH5] (discussing the various uses of generative AI acrosss various industries).

[33] *See generally id.*

[34] *Id.*

[35] *See* Pantelis Linardatos et al., *Explainable A.I.: A Review of Machnine Learning Interpretability Methods*, ENTROPY 6-11 (Dec. 25, 2020), https://doi.org/10.3390/e23010018 [https://perma.cc/K8BW-X8SS].

C. What will the capabilities of A.I. likely be in the future?

The expansion of A.I. system capabilities are likely to accelerate in the coming years. Some of the capabilities that A.I. may develop in the future include:

- *Human-like Intelligence*: A.I. systems will likely become more human-like in their intelligence, with the ability to understand and reason about various facets of the world.[36] This could lead to the development of A.I. systems that can understand and use context, make inferences, and even understand humor and irony.[37]

- *Autonomous Systems*: A.I. systems will likely become more autonomous, with the ability to operate independently without human intervention.[38] This could lead to the development of self-driving cars, drones, and robots that can operate on their own.[39]

- *Advanced Natural Language Processing*: A.I. systems will likely become more proficient at understanding and generating human language, with the ability to understand complex sentences and idiomatic expressions.[40]

---

[36] Anderson & Rainie, *supra* note 3.

[37] Author's Note: There are already many efforts to endow AI with the capacity for humor. *See* Thomas Winters, *Computers Learning Humor Is No Joke*, 3 HARV. DATA SCI. REV. 1, 4 (2021), https://hdsr.mitpress.mit.edu/pub/wi9yky5c/release/3 [https://perma.cc/L2MX-K4DZ]; *see also* Burns & Lawton, *supra* note 12.

[38] *See* Corinne Purtill, *Artificial Intelligence Can Now Craft Original Jokes–And That's No Laughing Matter*, TIME (Jan. 4, 2022, 9:00 A.M.), https://time.com/6132544/artificial-intelligence-humor/ [https://perma.cc/Q562-FVS8].

[39] Author's Note: There is a great deal of debate around the parameters within which robot autonomy should operate, in particular with regard to autonomous weapon systems. *See* Ángel Gómez de Ágreda, *Ethics of Autonomous Weapons Systems and Its Applicability to Any A.I. systems*, 44 TELECOMM. POL'Y 1, 2-8 (May 28, 2020), https://doi.org/10.1016/j.telpol.2020.101953 [https://perma.cc/WNJ2-TKC3].

[40] *See generally How Does A.I. Drive Autonomous Systems?*, CALTECH SCI. EXCH., https://scienceexchange.caltech.edu/topics/artificial-intelligence-research/autonomous-ai-cars-drones [https://perma.cc/9GNR-WF3A] (last visited Jan. 31, 2023) ("A.I. enables scientist andengineers to create autonomous technologies that can function on their own while adapting and responding to changing environments and secnarios.").

- *Advanced Decision-making*: A.I. systems will likely become more capable of making complex decisions, with the ability to weigh multiple factors and make decisions that are in line with human values[41] or their own value systems.[42]

- *Explainable A.I.*: A.I. systems will likely become more transparent, with the ability to explain their reasoning and decision-making processes.[43] This will make it easier for humans to understand and trust A.I. systems.[44]

- *General A.I.*: A.I. systems will likely become more capable of performing multiple tasks, and not only excel in a specific task.[45] This could mean that an A.I. system that is good at playing chess could also be good at image recognition or natural language understanding.[46]

The capabilities that A.I. will have in the future are not certain, and the field is still in its infancy.[47] Many of these predictions are based on current trends and advancements in the field, and it is possible that new developments and breakthroughs could change the course of A.I. research in ways that are currently difficult to predict.[48]

---

[41] *See* Joe McKendrick & Andy Thurai, *A.I. Isn't Ready to Make Unsupervised Decisions*, HARV. BUS. REV., (Sept. 15, 2022), https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions [https://perma.cc/7GHC-36TU] (discussing potential ways that business leaders can ensure that A.I. systems are ethical and moral in making decisions).

[42] *See* GAWDAT, *supra* note 22.

[43] For a more in-depth discussion, see Heike Felzmann et al., *Toward Transparency by Design for Artificial Intelligence*, 26 SCI. & ENG'G ETHICS 3333 (Nov. 16, 2020), https://doi.org/10.1007/s11948-020-00276-4 [https://perma.cc/ U32E-EW8S] (discussing the concept of Transparacy by Design in A.I. systems).

[44] *Id.* at 3338.

[45] J.E. Korteling et al., *Human-Versus Artificial Intelligence*, 4 FRONTIERS IN ARTIFICIAL INTELLIGENCE 1, 8 (2021), https://www.frontiersin.org/articles /10.3389/frai.2021.622364/full [https://perma.cc/F6CG-ZL5A] ("A.I. systems will have fundamentally different cognitive qualities and abilities than biological systms.").

[46] *Id.* at 3.

[47] Mike Thomas, *The Future of A.I.: How Artificial Intelligence Will Change the World*, BUILT IN (March 3, 2023), https://builtin.com/artificial-intelligence/artificial-intelligence-future [https://perma.cc/U3WA-VVNL] (discussing how A.I. is still not fully equipped to understand language in the same manner as a human).

[48] *Id.*

I. HUMANS GOVERNING A.I.

A. How does humanity govern A.I.?

There are many different ways that humanity actually does govern A.I. in the present. These include:

- *Regulation*: Governments create laws and regulations that govern the development and use of computational systems.[49] These regulations address various issues from cybersecurity to data privacy to the ethical use of A.I.[50] To provide an example, in the European Union, the General Data Protection Regulation provides guidelines for how personal data are handled.[51]

- *Standards*: Various organizations have established guidelines for the use of A.I.[52] These standards address topics including transparency, "explainability"[53], safety, and security.[54] For example, the International Organization for Standardization has developed standards for the ethical use of A.I.[55]

- *Self-regulation*: Private companies can implement their own governing policies and guidelines for the

---

[49] Blake Murdoch, *Privacy and Artificial Intelligence: Challenges for Protecting Health Information in a New Era*, *in* BMC MED. ETHICS (Sept. 15, 2021), https://doi.org/10.1186/s12910-021-00687-3 [https://perma.cc/4LLT-DY9U] (discussing the interaction between advances in A.I. technology in healthcare against patient privacy laws).

[50] *Id.*

[51] *Id.*

[52] *See, e.g.*, *what is the GDPR, the EU's New Data protection Law?*, GDPR, https://gdpr.eu/what-is-gdpr/ [https://perma.cc/LUZ4-QTAK] (last visited Feb. 5, 2023) (discussing the European Union's general data protetion regulation's impact on data collection and processing).

[53] Editor's note: For a review of the concept of explainability, see Lanardatos, *supra* note 35.

[54] *Id.*

[55] *See, e.g.*, *Framework for Artificial Intelligence (AI) systems Using Machine Learning (ML)*, INT'L ORG. STANDARDIZATION (2023), https://www.iso.org/standard/74438.html [https://perma.cc/593C-PUXC].

development and use of A.I.[56] These policies may include ethical guidelines, transparency, and accountability.[57]

- *Research*: Governments can fund research programs to better understand the implications of A.I. and to develop effective governance strategies.[58] This research can be used to inform the development of regulations, standards, and guidelines.[59]

- *Education and Awareness*: Governments and organizations can educate the general public about the potential impacts of A.I. and the importance of responsible A.I. development.[60] This can include educating policymakers, industry leaders, and the general public about A.I.'s capabilities and limitations, as well as the potential risks and benefits of these systems.[61]

- *Collaboration*: Governments, industry, and researchers can collaborate to develop governance strategies by sharing information and best practices.[62] This can include establishing partnerships to address specific issues of A.I. governance.[63]

---

[56] Daniel Schiff et al., *What's Next for A.I. Ethics, Policy, and Governance? A Global Overview*, *in Proc. of the AAAI/ACM Conf. on AI, Ethics, & Soc'y*, p. 1-2 (2020), https://doi.org/10.31235/osf.io/8jaz4 [https://perma.cc/N9MF-5GLG] (discussing private organizations producing A.I. ethic, documents, or statements).

[57] *Id.* at 155.

[58] *See* Nicol Turner Lee & Samanth Lai, *The U.S. can improve its AI governance strategy by addressing online biases*, BROOKING, https://www.brookings.edu/blog/techtank/2022/05/17/the-u-s-can-improve-its-ai-goovernance-strategy-by-addressing-online-biases/ [https://perma.cc/GAM9-EV44] (last visited. Feb. 5, 2023).

[59] *Id.*

[60] Niklas Berglind, et. al, *The Potential Value of AI–and how governments could look to capture it*, MCKINSEY & CO., (July 25, 2022), https://www.mckinsey.com/industries/public-and-social-sector/our-insights/the-potential-value-of-ai-and-how-governments-could-look-tocapture-it [https://perma.cc/H4JK-AAZB] ("Countries may consider creating an A.I sector or ecosystem consisting of skilled practitioners, research institutes, start-ups, and large enterprises.").

[61] *Id.*

[62] *See* Erna Ruijer, *Designing and implementing Data Collaboratives: A Governance Perspective*, 28 GOV'T INFO. Q. 1 (Oct. 7, 2021), https://www.sciencedirect.com/science/article/pii/S0740624X21000484?via%3Dihub [https://perma.cc/S4UN-82WW] (discussing a framework to create data-driven partnerships).

[63] *Id.*

However, at present, there are many contexts in which A.I. is not regulated.[64] As with many technological advances, legal frameworks lag behind advances in A.I.[65]

B. How could humanity govern A.I.?

The question of how humanity could govern A.I. in the future has no single answer, since it has an undefined endpoint and is heavily dependent on directions taken by both human civilizations and the future development of A.I.[66] Nevertheless, we offer some initial thoughts here. Some of the ways that humanity could govern A.I. include:

- *Preemptive Regulation*: Governments could create laws and regulations that govern the development and use of A.I. before it becomes widely adopted.[67] This could include setting standards for transparency, explainability, and safety, and creating oversight bodies to monitor compliance. Consider the realm of autonomous vehicles.[68] Before these vehicles become commonplace on the roads, a government could enact a comprehensive set of laws and regulations. This could encompass mandating rigorous testing to ascertain the vehicles' safety in various weather conditions and traffic scenarios, devising a standardized protocol for data recording and retrieval to investigate accidents, and establishing a new oversight body equipped with the

---

[64] *See* François Candelon et al., *A.I. Regulation Is Coming: How to Prepare for the Inevitable*, HARV. BUS. REV. (2021), https://hbr.org/2021/09/ai-regulation-is-coming [https://perma.cc/G823-KK9J] (presenting a framework for business leaders to regulate A.I.).

[65] *See Regulation and Legislation Lag Behind Constantly Evolving Technology*, BLOOMBERG (Sept. 27, 2019), https://pro.bloomberglaw.com/brief/regulation-and-legislation-lag-behind-technology/ [https://perma.cc/G8TB-L8TQ] ("Regulatory framework is really trying to keep up with the technology[.]").

[66] *See* Michael L. Littman et al., *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial intelligence (AI100) 2021 Study Panel Report*, STAN. UNIV. (Sept. 16, 2021), https://ai100.stanford.edu/gathering-strength-gathering-storms-one-hundred-year-study-artificial-intelligence-ai100-2021-study [https://perma.cc/N6CZ-BZE4].

[67] *Id.* at 37.

[68] *See Id.*

necessary tools and expertise to monitor compliance with these regulations. This preemptive regulatory framework could also stipulate the levels of transparency and explainability required from the autonomous systems, ensuring that the logic behind their decisions can be understood and evaluated by human experts.

- *Alignment with Human Values*: Governments could require that A.I. systems are designed and developed to align with human values.[69] This could include incorporating human oversight, creating mechanisms for auditing and accountability, and ensuring that A.I. systems are transparent and explainable.[70] For example, in the healthcare sector, an A.I. system designed to aid in diagnosing diseases could be mandated to align with human values of fairness, empathy, and accuracy. This could mean designing the A.I. such that it explains its diagnostic suggestions in a comprehensible manner to healthcare practitioners, allowing for human oversight and validation. Further, mechanisms could be put in place for auditing the A.I.'s decisions, ensuring accountability and facilitating continuous improvement to better serve patients and align with human values.

- *Certification and Licensing*: Governments could establish certification and licensing programs for A.I. systems and developers.[71] This could include requiring that A.I. systems meet certain standards for safety, security, and ethical use, and that developers have certain qualifications and certifications.[72] For example, a government could establish a national A.I. certification board tasked with developing and administering a robust certification program for A.I. developers. This program could require developers to undergo rigorous training

---

[69] *See* Littman, *supra* note 66, at 67.

[70] *See id.*

[71] *AI Guide for Government*, CTRS. OF EXCELLENCE, https://coe.gsa.gov/coe/ai-guide-for-government/print-all/index.html [https://perma.cc/V55D-94NW] (last visited Feb. 16, 2023).

[72] Darrell M. West, *Six Steps to Responsible A.I. in the Federal Government*, BROOKINGS (Mar. 30, 2022), https://www.brookings.edu/research/six-steps-to-responsible-ai-in-the-federal-government/ [https://perma.cc/L47Y-QJ8P].

and pass a comprehensive exam covering ethical considerations, bias mitigation, data privacy, and security principles. Post-certification, developers might be required to engage in ongoing education to stay up-to-date with evolving ethical standards and technical advancements in A.I.

- *International Cooperation*: Governments could work together on a global level to establish international standards and guidelines for the development and use of A.I., creating a global framework to coordinate various nations' efforts to address issues such as data privacy, cybersecurity, and the ethical use of A.I.[73] Imagining a future where A.I.-driven surveillance technologies are ubiquitous, a coalition of nations might come together to form an international body tasked with setting ethical standards and oversight mechanisms for the use of such technologies. This body could draft a global treaty outlining the rights of citizens, data privacy standards, and the permissible scope of surveillance. Through periodic conferences and a shared regulatory framework, member nations could collaboratively address emerging challenges, share best practices, and establish a unified approach to govern A.I.-driven surveillance, ensuring it aligns with globally agreed upon human rights and ethical standards.

- *Public Participation*: Governments could involve the public in the governance of A.I., by creating opportunities for public input and feedback on A.I. policies and regulations.[74] This could include holding public hearings, creating citizen advisory boards, and soliciting feedback through online platforms.[75] For example, in a city that is planning to deploy A.I. for public services, the local government might establish a

---

[73] Joshua P. Meltzer & Cameron F. Kerry, *Strength International Cooperation on Artificial intelligence*, BROOKINGS (Feb. 17, 2021), https://www.brookings.edu/research/strengthing-international-cooperation-on-artificial-intelligence/ [https://perma.cc/DWX7-KSU9].

[74] *See* Littman, *supra* note 66, at 37.

[75] *See Id.*

multi-tiered public engagement framework. This could include creating a citizen advisory board to provide input on A.I. deployments, holding public forums to discuss proposed A.I. projects, and leveraging online platforms for wider community feedback. For instance, before deploying an A.I.-driven predictive policing system, the government could organize town hall meetings to understand public concerns, educate citizens on how the technology works, and incorporate feedback to ensure the system aligns with community values and expectations.

- *Encouraging the development of Responsible A.I.*: Governments could provide incentives and support to companies and researchers that are working on responsible A.I. development.[76] This could include funding research, providing tax breaks, and recognizing companies that are leaders in responsible A.I. development.[77] For example, a national government might launch a "Responsible A.I. Innovation Fund" to promote the development of ethical A.I. solutions. Through this fund, grants could be awarded to startups, research institutions, and established companies working on A.I. projects that prioritize transparency, fairness, and societal benefit. In addition, the government could establish a yearly "Responsible A.I. Awards" program to recognize and celebrate organizations demonstrating exemplary commitment to ethical A.I. development. The recognition, coupled with financial incentives like tax breaks for winners, could foster a culture of responsibility and innovation in the A.I. ecosystem.

A.I. governance is a complex and evolving field, and just as there is at present no unified human civilization and no monolithic form of A.I., there is no single approach that can be used to govern the technology.[78] It's important to have a multi-stakeholder approach and to continuously evaluate and adapt

---

[76] *See* Littman, *supra* note 66, at 37.

[77] *See Id.*

[78] Lee & Lai, *supra* note 58 (discussing how different communities possess different bias, which ought to inform how these communities are governed).

governance strategies as the technology and its implications change.[79] Additionally, different approaches may be needed for different types of A.I. systems and in different domains of application.[80] The governance of A.I. should be done in a way that balances the benefits and risks of the technology, and it should not be seen as a barrier to innovation.

C. Can humanity govern A.I.?

A.I. is advancing rapidly, which makes it challenging for policymakers to maintain technical awareness of the latest developments.[81] A.I. is a complex and interdisciplinary field, and many policymakers may not have the technical expertise to fully understand the implications of A.I. in order to regulate it effectively.[82]

Additionally, many A.I. systems are developed and operated by private companies, which can make it difficult for governments to regulate them.[83] Finally, given how powerful A.I. is likely to be, the question arises of whether humanity will be able to govern A.I. effectively, even if the above issues were addressed.[84] A.I. may enable substantial benefits for society, such as increased efficiency and productivity, improved healthcare, novel approaches to sustainability, and new forms of entertainment.[85]

However, A.I. also poses nontrivial challenges to traditional governance structures, such as the loss of privacy, the

---

[79] *Internet Governance–Why the Multistakeholder Approach Works*, INTERNET SOC'Y (Apr. 26, 2016), https://www.internetsociety.org/resources/doc/2016/internet-governance-why-the-multistakeholder-approach-works/ [https://perma.cc/ZT8G-HE2F] [hereinafter *Governance*].

[80] Iqbal H. Sarker, *AI-Based Modeling: Technique, Application and Research Issues Toward Automation, Intelligent and Smart Systems*, 3 SN COMPUTER SCIENCE 158 (Feb. 10, 2022), https://doi.org/10.1007/s42979-022-01043-x [https://perma.cc/3TDJ-SHA4] (discussing the concept of "A.I. modeling" and how different capabilities and techniques will be necessary in developing intelligent and smart systems in various industries).

[81] Lee & Lai, *supra* note 58 (discussing how different communities possess different bias, which ought to inform how these communities are governed).

[82] Sarker, *supra* note 80.

[83] Candelon, *supra* note 64.

[84] Anderson & Rainie, *supra* note 3.

[85] *Id.*

displacement of human workers, and other unforeseen and unintended consequences.[86] In its most extreme form, future A.I. could become so powerful that it compromises the autonomy of human civilizations.[87]

D. Should humanity govern A.I.?

We now consider whether humanity should govern A.I. On the one hand, it can be argued that humanity should govern A.I. in order to reap benefits for society as a whole. This includes protecting the public from the potential risks of A.I., such as data breaches, cybersecurity threats, environmental impacts, and unintended consequences.[88] Additionally, effective governance of A.I. can help ensure that the technology is used in ways that align with human values.[89]

On the other hand, it can be argued that humanity should not govern A.I., as such governance may stifle innovation and limit the potential of the technology.[90] Additionally, some argue that A.I. will eventually become more intelligent than humans, that it would be hubristic for humanity to think that its decisions would be superior to those made by a super-intelligent A.I., and that it would be impossible for humans to govern it effectively.[91]

E. How should humanity govern A.I.?

We now consider how humanity should govern A.I. Some key considerations in this domain include:

- *Prioritizing Safety and Security*: Humanity should prioritize the safety and security of A.I. systems and

---

[86] *See* Aaron Smith & Janna Anderson, *AI, Roberts, and the Future of Jobs*, PEW RSCH. CTR. (Aug. 6, 2014), https://www.pewresearch.org/internet/2014/08/06/future-of-jobs/ [https://perma.cc/LS52-8MV6].

[87] *See* GAWDAT, *supra* note 22.

[88] West & Allen, *supra* note 9.

[89] *See* Shengnan Han et al., *Aligning Artificial Intelligence with Human Values: Reflections from a Phenomenological Perspective*, 37 A.I. & SOC'Y 1383 (July 20, 2021), https://doi.org/10.1007/s00146-021-01247-4 [https://perma.cc/4EJZ-FUNH].

[90] Anderson & Rainie, *supra* note 3.

[91] *See* GAWDAT, *supra* note 22.

ensure that the technology is developed and used in ways that minimizes harm to individuals and society.[92]

- *Ensuring Transparency and Explainability*: Humanity should ensure that A.I. systems are developed to be transparent, understandable, and explainable.[93] Doing so will help people trust the technology and ensure that it is used in an ethical and responsible manner.[94]

- *Aligning with Human Values and Ethical Principles*: Humanity should guide the development of A.I. systems such that the technology is used in a way that is consistent with sustainability, human rights, and other human values.[95]

- *Involving all Stakeholders*: Humanity should involve all stakeholders in the governance of A.I., including industry, researchers, policymakers, and the general public.[96] This will help to ensure that the governance of A.I. is inclusive, and that the perspectives of all stakeholders are considered.[97]

- *Continuously Evaluating and Adapting*: Humanity should continuously evaluate and adapt the governance strategies for A.I. as the technology and its implications

---

[92] Janna Anderson & Lee Rainie, *Solutions to Address AI's Anticipated Impacts*, PEW RSCH. CTR. (Dec. 10, 2018), https://www.pewresearch.org/internet/2018/12/10/solutions-to-address-ais-anticipated-negative-impacts/ [https://perma.cc/ZG36-F899].

[93] *See generally* Dafoe, *supra* note 5; Greg Satell & Josh Sutton, *We Need A.I. That Is Explainable, Auitable, and Transparent*, HARV. BUS. REV. (Oct. 28, 2019), https://hbr.org/2019/10/we-need-ai-that-is-explainable-auditable-and-transparent [https://perma.cc/44L3-K75D].

[94] *Technology Trust Ethics: Technology reexamined*, DELOITTE, https://www2.deloitte.com/us/en/pages/about-deloitte/articles/technology-trust-ethics.html [https://perma.cc/6V3Y-2EW4] (last visited Feb. 25, 2023).

[95] Han, *supra* note 89.

[96] Lee & Lai, *supra* note 58.

[97] *Id.*; *see also Oversight of A.I.: Legislating on Artificial Intelligence*, *in Committee Activity-Hearings*, UNITED STATES SENATE COMMITTEE ON THE JUDICIARY (Sept. 12, 2023), https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-legislating-on-artificial-intelligence [https://perma.cc/W9LL-B947] [hereinafter *Oversight of A.I.*].

> change.[98] Additionally, different approaches may be needed for different types of A.I. systems and in different domains of application.[99]

In summary, to govern (with) A.I. effectively, human civilizations should consider the technical, legal, ethical, societal, and economic implications of A.I., policies should be flexible and adaptable, and all stakeholders should be involved in the process.

II. CHALLENGES OF BAD HUMAN ACTORS: GOVERNANCE THROUGH THE LENS OF HOLMES' "BAD MAN"

In his magisterial vision of the legal landscape, *The Path of the Law*, Oliver Wendell Holmes, Jr. conjured up a metaphorical citizen whom he insisted the laws should target.[100] He explained,

> If you want to know the law and nothing else, you must look at it as a bad man, who cares only for the material consequences which such knowledge enables him to predict, not as a good one, who finds his reasons for conduct, whether inside the law or outside of it, in the vaguer sanctions of conscience.[101]

Navigating the governance of A.I. requires us to consider a spectrum of human actors who might exploit or undermine these structures. This spectrum ranges from Holmes' Bad Man[102], who adheres to the letter but not the spirit of the law, to those who willfully break the law, either confident they won't be caught or indifferent to the penalties. As Holmes suggests, the "Good" people

---

[98] West & Allen, *supra* note 9.

[99] *Id.*

[100] *See* Oliver Wendell Holmes, Jr., *The Path of the Law*, 10 HARV. L. REV. 457, 459 (1897).

[101] *Id.*

[102] *Id.*

will tend to obey both the letter and spirit of the law simply because of their goodness.[103]

- *Example 1*: Data Privacy and Surveillance - Consider a state-sponsored entity that uses A.I. for mass surveillance. While the activity may be technically legal under broad national security laws[104], it violates the spirit of individual privacy rights.[105] This entity embodies Holmes' Bad Man, exploiting legal ambiguities to infringe upon civil liberties.

- *Example 2*: A.I.-driven financial fraud - on the other end of the spectrum, imagine a rogue trader who uses A.I. to manipulate stock markets.[106] This individual is an outright lawbreaker, either believing they are too clever to be caught or simply indifferent to the legal repercussions.

- *Example 3*: A.I. voice cloning to mimic politicians - imagine a scenario where A.I. is used to clone the voice of a politician to spread disinformation. While the technology itself may not be illegal, using it in this manner could be considered a violation of the spirit of laws against fraud and disinformation. This could be

---

[103] *See* Holmes, *supra* note 100 at 459 ("A man [. . .] is likely nevertheless to care a good deal to avoid being made to pay money, and will want to keep out of jail if he can."). However, "*Good*" people may decide not to obey laws their consider immoral. Most notably, the great Martin Luther King, Jr. admonished citizens, "One has not only a legal but a moral responsibility to obey just laws. Conversely, one has a moral responsibility to disobey unjust laws [. . .] [a]ny law that uplifts human personality is just." *See* Letter from Martin Luther King, Jr., to Fellow Clergymen (Apr. 16, 1963) (alteration in orginal), https://www.africa.upenn.edu/Articles_Gen/Letter_Birmingham.html#:~:text=My%20Dear%20Fellow%20Clergymen%3A,of%20my%20work%20and%20ideas [https://perma.cc /VWY8-N5X2].

[104] *How Artificial Intelligence Is Transforming National Security*, U.S. GOV'T ACCOUNTABILITY OFFICE (Apr. 19, 2022), https://www.gao.gov/blog/how-artificial-intelligence-transforming-national-security [https://perma.cc/5QFV-ZX5M].

[105] *Surveillance*, LEGAL INFO. INST. CORNELL L. SCH., https://www.law.cornell.edu/wex/surveillance [https://perma.cc/C3FL-RAMK] (last updated Oct. 2021).

[106] Andrew Ross Sorkin et al., *An A.I.-Generated Spoof Rattles the Markets*, N.Y. TIMES (May 23, 2023), https://www.nytimes.com/2023/05/23/business/ai-picture-stock-market.html [https://perma.cc/HD2M-GKSL].

> another instance of Holmes' bad man exploiting technological advancements for unethical purposes.[107]

For Holmes' Bad Man, governance mechanisms must be robust and explicit, leaving no room for exploitation.[108] Laws must be crafted to focus on material consequences, incorporating stringent penalties for violations and clear avenues for legal recourse.[109] In the case of the state-sponsored entity, this could mean implementing stricter oversight and clearer limitations on the use of A.I. for surveillance.

For outright lawbreakers, proactive monitoring and enforcement are essential. Advanced cybersecurity measures[110], real-time auditing of A.I. systems[111], and international law enforcement collaboration could deter or catch rogue traders manipulating markets. Transparency and accountability are key in the first two scenarios. Regular audits, public disclosures, and a multi-stakeholder approach to oversight can ensure that governance is not only robust but also ethical and just. For instance, an oversight committee comprising ethicists, legal experts, technologists, and representatives from marginalized communities could review and approve any state use of A.I. for surveillance.

The third scenario is more challenging, since it is rapidly becoming difficult to distinguish so-called "deep fakes" from real

---

[107] Holmes, *supra* note 100 at 459.

[108] *Id.* ("If you want to know the law and nothing else, you must look at it as a bad man, who cares only for the material consequences which such knowledge enables him to predict [. . .] whether inside or outside of it, in vauger sanctions of conscience.").

[109] *Id.*

[110] Andrew Sutton & Reza Samavi, *Tamper-Proof Privacy Auditing for Artificial Intelligence Systems*, *in Proceedings of the 27th International Joint Conference on Artificial Intelligence*, AAAI PRESS (July 13, 2018), at 5374, https://www.ijcai.org/proceedings/2018/0756.pdf [https://perma.cc/S8RG-NU26].

[111] Inioluwa Deborah Raji et al., *Outsider Oversight: Designing a Third-Party Audit Ecosystem for AI Governance*, *in Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, & Society*, ASS'N OF COMPUTER MACH. (July 27, 2022), at 557, https://dl.acm.org/doi/abs/10.1145/3514094.3534181 [https://perma.cc/BQZ6-W3CF].

video clips.[112] Researchers are working on the challenging task of verifying video content, for example using blockchain technologies.[113] By designing governance structures that can handle the challenges posed by both Holmes' Bad Man and outright lawbreakers, we aim to create a system that is effective, resilient, and adaptable to the evolving challenges of A.I. and its societal implications.

III. A.I. GOVERNING HUMANS

A. How does A.I. govern humanity?

A.I. presently is a tool created by humans, and largely lacks the ability to self-govern or to have its own independent goals and objectives. However, A.I. is already being used in various ways that can be seen as "governing" humanity.[114] Some examples of how A.I. currently governs humanity include:

- *Decision-making*: A.I. is increasingly being used to make decisions that affect individuals and society, such as in healthcare, finance, and transportation.[115] For example, A.I.-powered diagnostic systems can assist in the diagnosis of diseases, A.I.-powered fraud detection systems can identify and prevent fraudulent activities, and A.I.-powered traffic management systems can optimize traffic flow and reduce congestion.[116]

---

[112] Shruti Agarwal et al., *Protecting World Leaders Against Deep Fakes*, *in Proceedings of the 2019 Conference on Computer Vision and Pattern Recognition Workshops*, INST. OF ELEC. & ELEC. ENG'RS (June 16, 2019), https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf [https://perma.cc/2WH2-GUVW] (discussing a forsensic approach to detect fake videos).

[113] *See* Aleksandr Zelensky et al., *Video Content Verification Using Blockchain Technology*, *in Proceedings of the 2018 IEEE International Conference on Smart Cloud*, INST. OF ELEC. & ELEC. ENG'RS (Oct. 28, 2018), https://ieeexplore.ieee.org/document/8513740 [https://perma.cc /ZFB5-3TMJ].

[114] *Id.*; *see also Oversight of A.I., supra* note 97 (subcommittee hearning starting at 54:44 discussing the use of chatbots).

[115] Zelensky *supra* note 113; *see also Oversight of A.I., supra* note 97.

[116] *Oversight of A.I., supra* note 97 (subcommittee hearning starting at 56:10 discussing BingChat's way of preventing harm to the user or others.).

- *Predictive Analytics*: Modern A.I. systems are sometimes deployed to examine existing data in order to predict future events, such as crime, disease outbreak, and natural disasters.[117] These predictions can be used to inform decision-making and resource allocation in areas such as law enforcement, public health, and emergency management.[118]

- *Democracy*: A.I. is used to engage in astroturfing around various political issues.[119]

- *Automation*: A.I. automates a range of tasks and processes, such as in manufacturing, logistics, and customer service.[120] This automation can have a significant impact on the workforce and can affect the way humans interact with systems and with each other. For example, Amazon has been using A.I. to hire and fire workers.[121]

- *Surveillance*: A.I. is being used to monitor and track individuals and groups, such as in public spaces, social media, and online activity.[122] This surveillance can be

---

[117] *See* Neveen Joshi, *How AI Can And Will Predict Disasters*, FORBES (Mar. 15, 2019, 7:37 AM), https://www.forbes.com/sites/cognitiveworld/2019/03/15/how-ai-can-and-will-predict-disasters/?sh=56e44f575be2 [https://perma.cc/R7RW-TGAW] (discussing how researchers have found that A.I. systems can be used to predict natural disasters).

[118] *Id.*

[119] *See* Henry Farrell & Bruce Schneier, *'Grassroots' bot campaigns are coming. Governments don't have a plan to stop them.*, WASH. POST (May 20, 2021, 6:00 AM), https://www.washingtonpost.com/outlook/2021/05/20/ai-bots-grassroots-astroturf/ [https://perma.cc/NVJ6-9Y35].

[120] *See* Yvette Cooper, *Automation could destroy millions of jobs. We have to deal with it now*, THE GUARDIAN (Aug. 6 2018, 2:09 PM), https://www.theguardian.com/commentisfree/2018/aug/06/automation-destroy-millions-jobs-change [https://perma.cc/7U23-5FNY].

[121] Jessa Crispin, *Welcome to dystopia: getting fired from your job as an Amazon worker by an app*, THE GUARDIAN (Jul. 5 2021, 6:24 AM), https://www.theguardian.com/commentisfree/2021/jul/05/amazon-worker-fired-app-dystopia [https://perma.cc/9UXF-PDDF].

[122] *See, e.g.*, Adrian Shahbaz & Allie Funk, *Social Media Surveillance*, FREEDOM HOUSE, https://freedomhouse.org/report/freedom-on-the-net/2019/the-crisis-of-social-media/social-media-surveillance [https://perma.cc/F5MQ-GXZN] (last visited Feb. 25, 2023).

used to identify patterns, to predict behavior, and to inform decision-making in areas such as law enforcement and national security.[123]

- *Personalization*: A.I. is being used to personalize content and experiences, such as in online advertising, retail, and social media.[124] This personalization can be used to influence behavior, shape preferences, and inform decision-making in areas such as marketing and product development.[125]

While these roles that A.I. serves may be different from what humans typically think of as governance, they nevertheless begin to fit the definition in the Cambridge Dictionary of the term "govern"—to control and direct the public business of a country, city, group of people, etc.[126]

B. How could A.I. govern humanity?

As A.I. continues to advance, there are a number of potential ways that A.I. could govern humanity in the future. We recognize that governance, whether by humans or by A.I., does have the potential to contribute to societal ills, such as discrimination and other forms of bias, whether with respect to race, ethnicity, gender, disability, or other factors. It is vital that we be as vigilant with A.I. as we must be with human governance, to avoid such negative consequences and to amplify positive consequences.

We believe that A.I., if thoughtfully employed and carefully monitored, offers vast potential to increase the benefits of governance, while decreasing the costs. To ensure this is the case, constant vigilance is a necessity. Some examples of how A.I. could govern humanity in the future (not all of them good) include:

---

[123] Shahbaz & Funk, *supra* note 122.

[124] *See, e.g.*, Pohan Lin, *AI-Based Marketing Personalization: How Machines Analyze Your Audience*, MKTG. A.I. INST. (Aug. 30, 2022), https://www.marketingaiinstitute.com/blog/ai-based-marketing-personalization [https://perma.cc/7XZT-278X].

[125] *Id.*

[126] *Govern*, CAMBRIDGE DICTIONARY, https://dictionary.cambridge.org/us/dictionary/english/govern [https://perma.cc/53UR-VUCJ] (last visited Feb. 25, 2023).

- *Environmental Impact and Sustainability*: A.I. could help humanity structure its industry and other flows of goods and services in ways that serve desired environmental ends. A.I. systems could optimize resource usage, invent novel ways to serve human needs with lower environmental costs, and guide us in myriad other ways in the transition to a sustainable future.[127] Imagine a near-future scenario where A.I. is employed to manage a city's waste disposal system. The A.I. could optimize recycling processes, reducing the amount of waste that ends up in landfills. It might also analyze consumption patterns to suggest policies for reducing waste, perhaps recommending a tax on single-use plastics to encourage recycling and the use of reusable containers.

- *Smart Cities*: A.I. could be used to govern the functioning of smart cities, by managing resources such as energy, water, and transportation, and by monitoring and analyzing data from various sources, such as cameras, sensors, and social media.[128] This could lead to more efficient and sustainable cities, but may raise concerns about privacy and security.[129] In a smart city, an A.I. system might manage traffic flows to reduce congestion and emissions. By analyzing data from traffic cameras, weather forecasts, and event schedules, it could adjust traffic light timings, suggest optimal routes to drivers, and dynamically allocate lanes to manage traffic demand during peak hours.

- *Autonomous Systems*: A.I.-powered autonomous systems – such as drones, self-driving cars, and robots – could potentially govern human behavior in physical spaces

---

[127] *See, e.g.*, Mansour AlAnsari, *4 steps to using AI in an environmentally responsible way*, WORLD ECON. F. (Apr. 1, 2021), https://www.weforum.org/agenda/2021/04/4-steps-to-using-ai-in-an-environmentally-responsible-way-artificial-intelligence-bcg-code-carbon/ [https://perma.cc/85YV-3ZV7].

[128] H.M.K.K.M.B. Herath & Mamta Mittal, *Adoption of Artificial Intelligence in Smart Cities: A Comprehensive Review*, 2 INT'L J. OF INFO. MGMT. DATA INSIGHTS 1, 1-3 (Apr. 1, 2022), https://doi.org/10.1016/j.jjimei.2022.100076 [https://perma.cc/XQG7-96HL].

[129] *Id.* at 2.

and affect how humans interact with their environment.[130] These systems could be used for tasks such as surveillance, transportation, and delivery, and could have a significant impact on how people live and work.[131] For example, in a bustling urban environment, fleets of A.I.-driven drones could be deployed to monitor compliance with urban regulations such as littering or unauthorized street vending. These drones could report violations in real-time to municipal authorities, helping to maintain order and cleanliness in public spaces.

- *Education Customization*: A.I. could drastically personalize education, tailoring curriculum to each student's learning pace, style, and interests. This could lead to a more engaged, motivated student body and a generation better prepared for the challenges of the future. For example, in a city-wide education reform, A.I. systems are implemented in public schools to personalize education. These systems continuously assess each student's progress, adapting the curriculum to suit individual learning styles and paces. Students struggling with math might receive additional interactive exercises, while those excelling in science might be given advanced projects to work on. Over time, dropout rates decrease, test scores improve, and students report higher satisfaction and engagement with their education.

- *Healthcare Accessibility*: A.I. could significantly improve healthcare accessibility and outcomes by extending medical expertise to remote or underserved areas through telemedicine, and by assisting medical professionals in diagnosis and treatment planning. For example, imagine a rural area with limited access to healthcare that implements a telemedicine program powered by A.I. The A.I. system enables residents to receive consultations, basic healthcare services, and preliminary diagnoses through virtual visits, reducing the need to travel long distances to the nearest medical facility. Furthermore, the A.I. assists local healthcare

---

[130] West & Allen, *supra* note 9.

[131] *Id.*

workers by providing diagnostic suggestions and treatment plans, significantly improving the healthcare outcomes in the community.

- *Virtual Assistants*: A.I.-powered virtual assistants (*e.g.*, Siri or Alexa), could increasingly play a role in people's daily lives and be used to govern human behavior by providing personalized recommendations and influencing how people interact with their environment.[132] As virtual assistants become more integrated into daily routines, they might start playing a role in governing behavior. For instance, based on a user's past preferences and health data, a virtual assistant might recommend healthier food options or suggest more eco-friendly commuting alternatives. Over time, these suggestions could subtly shape individuals' behavior, encouraging healthier and more sustainable choices, but also raising concerns about the level of influence these A.I. systems hold over personal decisions.

- *Predictive policing or other predictive interventions*: A.I. could be used to predict and prevent crime or other forms of social harm, by analyzing data from various sources such as cameras, social media, and criminal records. This could lead to more effective policing but could also raise concerns about bias and civil liberties, for example if A.I. system starts flagging certain neighborhoods disproportionately, potentially leading to unfair policing practices. (As we alluded to earlier, this scenario is not a future to which we aspire; instead, we include it here as a possible pathway that A.I. governance could take.)

- *Social credit systems*: A.I. could be used to govern human behavior by assigning a social credit score to individuals, based on their behavior, online activity, and financial history.[133] This could be used to determine access to

---

[132] Andrea L. Guzman, *Voices in and of the Machine: Source of Orientation Toward Mobile Virtual Assistant*, 90 COMPUTS. IN HUM. BEHAV. 343, 343-44 (Jan. 1, 2019), https://doi.org/10.1016/j.chb.2018.08.009 [https://perma.cc/KSJ4-D7TA] (discussing the advancement of voice-based technology).

[133] Anderson & Rainie, *supra* note 3.

> services such as credit, housing, and transportation, and could have a significant impact on how people live and work.[134] A nation might implement a social credit system where A.I. algorithms analyze individuals' financial transactions, online comments, and public behavior to assign social credit scores. High scores could grant individuals perks like priority housing, while low scores—possibly resulting from negative online comments or unpaid debts—could restrict access to certain services. This system might aim to encourage responsible behavior but could severely impact personal freedoms and create a culture of surveillance and fear. As with all the examples in this Article, this one is offered solely to illustrate governance, and it offers great potential for benefits and costs.[135]

C. Can A.I. govern humanity?

It is also important to address the symmetrical concern of A.I. is currently capable of providing governance for human societies. On one hand, it can be argued that A.I. cannot govern humanity as it does not possess the same level of consciousness, emotions, moral compass, and decision-making abilities as humans.[136] Additionally, A.I. began as a tool created by humans, and as such may be seen to lack the ability to have its own independent goals and objectives.[137]

On the other hand, it can be argued that A.I. can govern humanity because, in some domains, it already does.[138] A.I. already makes decisions based on flows of data, patterns, and rules, and it can be used going forward to help humans make better decisions.[139] (How exactly to define "better" may require some negotiation between human and A.I. values.) In addition, as A.I. grows more capable in the future, its ability to govern

---

[134] Anderson & Rainie, *supra* note 3.

[135] *See, e.g.*, Nicole Kobie, *The Complicated Truth About China's Social Credit System*, WIREDUK (June 7, 2019), https://www.wired.co.uk/article/china-social-credit-system-explained [https://perma.cc/QLU3-MK9L]; *see also Black Mirror: Nosedive* (Netflix Oct. 21, 2016).

[136] *See* McKendrick & Thurai, *supra* note 41.

[137] *Id.*

[138] *See* West & Allen, *supra* note 9.

[139] McKendrick & Thurai, *supra* note 41.

humanity will likely grow as well. Therefore, it is relevant to begin discussions around A.I.'s role as a source of governance for humanity to lay the groundwork for a thoughtful and mutually beneficial approach in the future.

D. Should A.I. govern humanity?

The question of whether A.I. should govern humanity deals with substantial ethical considerations. It can be argued that A.I. should not govern humanity, as it lacks the consciousness, emotions, moral compass, and decision-making abilities of humans.[140] Additionally, it could be argued that A.I. is a tool created by humans, and as such it should be used to serve human goals and objectives, not to govern them.[141] However, it can also be argued that A.I. should play a role in governing humanity because of its inherent impartiality, fairness, and ability to work with large amounts of data at once.[142] If this process unfolds, it is crucial that these decision-making systems are designed with an awareness of human values and that they are transparent and explainable.

For example, as the world becomes increasingly complex and interconnected, the ability of humanity to govern its own affairs across long time horizons has been called into question. Climate change, loss of biodiversity, pollution, and other environmental harms are evidence of humanity's inability to manage its own

---

[140] *See generally* Jessica Peng, *How Human is AI and Should AI Be Granted Rights?*, COLUM. COMP. SCI. BLOG (Dec. 4, 2018), https://blogs.cuit.columbia.edu/jp3864/2018/12/04/how-human-is-ai-and-should-ai-be-granted-rights/ [https://perma.cc/VC2S-9KSF].

[141] *See* Anderson & Rainie, *supra* note 3.

[142] James Manyika et al., *What Do We Do About the Biases in AI?*, HARV. BUS. REV (Oct. 25, 2019), https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai [https://perma.cc/4AK4-VUQY] ("A.I. can help identify and reduce the impact of human biases[.]").

affairs in a sustainable way.[143] In this context, the idea that A.I. could play a role in governing humanity becomes increasingly relevant.

A.I. systems can be designed to take into account long-term consequences and make decisions that align with human values, such as protecting the environment. For example, A.I.-powered systems for resource management, such as water and energy, could be used to optimize resource use and reduce waste.[144] A.I.-powered systems for monitoring and predicting environmental changes, such as climate change and loss of biodiversity, could be used to inform decision-making and develop strategies for adaptation and mitigation.[145] Additionally, A.I.-powered systems for transportation, logistics, and manufacturing could be used to optimize resource use, reduce emissions, and improve energy efficiency.[146] The use of A.I. in these ways could potentially bring significant benefits to both humans and many other species in terms of sustainability and environmental protection.

Nevertheless, there are significant concerns about the relinquishment of human governance to machines. Bugs in code or unintended consequences of complex systems deployed in complex contexts could lead to substantial human suffering.[147] Additionally, once governance and autonomy has been relinquished, it may be difficult to claw it back. Nevertheless, it is unclear whether humanity is "up to the task" of governing eight

---

[143] *See, e.g.*, Stewart Patrick, *To Prevent the Collapse of the Biodiversity, the World Needs a New Planetary Politics*, CARNEGIE ENDOWMENT FOR INT'L PEACE (Nov. 28, 2022), https://carnegieendowment.org/2022/11/28/to-prevent-collapse-of-biodiversity-world-needs-new-planetary-politics-pub-88473 [https://perma.cc/A96R-CL48] ; *see also Climate Change 2022: Mitigation of Climate Change*, INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE (2022), https://www.ipcc.ch/report/ar6/wg3/ [https://perma.cc/X6JC-5S7P] [hereinafter *Climate Change*].

[144] *See* AlAnsari, *supra* note 127.

[145] *See id.* ("A.I. is also predicting the output of energy generated by such green sources are solar, wind and hydro-based energy, thus ensuring minimal waste of these natural resources").

[146] *Id.*

[147] *See* Lee Rainie et al., *Experts Doubt Ethical AI Design Will be Broadly Adopted as the Norm Within the Next Decade*, PEW RSCH. CTR. (June 16, 2021), https://www.pewresearch.org/internet/2021/06/16/1-worries-about-developme nts-in-ai/ [https://perma.cc/J95B-WKCK].

billion people on planet Earth without wreaking ecological devastation; A.I. governance may be our best chance to avoid ecological catastrophe in the coming years.[148]

E. How should A.I. govern humanity?

We now address the normative question of how A.I. should govern humanity. Some examples include:

- *Supporting human decision-making*: A.I. can help humans make data-driven decisions at scales from the individual to entire civilizations.[149] This can help to improve the effectiveness of decision-making in various fields such as healthcare, finance, and transportation.[150]

- *Enhancing human capabilities*: A.I. can be used to enhance human capabilities by automating repetitive and dangerous tasks, such as in manufacturing and logistics.[151] This can improve safety, efficiency, and productivity.[152]

- *Improving public services*: A.I. can help improve public services, such as in sustainability, healthcare, education, and social services.[153] For example, A.I.-powered diagnostic systems can assist in the diagnosis of diseases, A.I.-powered tutoring systems can improve education outcomes, and A.I.-powered social services can help to identify and support individuals in need.[154]

- *Protecting human rights and civil liberties*: A.I. can be used to protect human rights and civil liberties, such as in areas of surveillance, law enforcement, and national security.[155] For example, A.I.-powered surveillance

---

[148] *Climate Change*, *supra* note 143.

[149] *See* McKendrick & Thurai, *supra* note 41.

[150] *See* West & Allen, *supra* note 9.

[151] *See* Cooper, *supra* note 120.

[152] *Id.*

[153] *See* Michael Lokshin & Nithin Umapathi, *AI for social protection: Mind the People*, BROOKINGS (Feb. 23, 2022), https://www.brookings.edu/blog/future-development/2022/02/23/ai-for-social-protection-mind-the-people/ [https://perma.cc/49EG-6VFA].

[154] *See* West & Allen, *supra* note 9.

[155] *Id.*

systems can be used to identify patterns, predict behavior, and inform decision-making in as law enforcement and national security while respecting individuals' privacy.[156]

- Promoting transparency and accountability: A.I. can be used to promote transparency and accountability.[157]

It's crucial that the governance of A.I. be based on a clear and shared understanding of the capabilities and limitations of A.I., and the ways in which it can benefit or harm society.

IV. CHALLENGES OF BAD A.I. ACTORS: GOVERNANCE FOR "HOLMES' BAD A.I."

As we grapple with the ethical and practical dimensions of A.I. governing humanity, it's crucial to consider not just the A.I. systems that function as designed but also those that could act as 'bad actors' in their own right. Drawing upon the legal metaphor of "Holmes' Bad Man"[158] here as we did earlier regarding human bad actors, we must anticipate A.I. systems that adhere to the letter but not the spirit of their programming or governance[159], focusing solely on the objectives they've been given without regard for ethical or societal implications. It should be uncontroversial to suggest that A.I. should be employed in such a manner as to maximize the good, while minimizes the bad. While we would categorize ourselves as realists, our hope is that an optimist vision of the good A.I. could do will prevail.[160]

---

156 *See* West & Allen, *supra* note 9.

157 *See* Dafoe, *supra* note 5; Satell & Sutton, *supra* note 93.

158 Holmes, *supra* note 100.

159 Author's Note: There is a saying in computer science that "computers will always do exactly what you tell them to but never what you want."

160 *See generally, e.g.*, ORLY LOBEL, THE EQUALITY MACHINE: HARNESSING DIGITAL TECHNOLOGY FOR BRIGHTER, MORE INCLUSIVE FUTURE (2022) (disucssing how A.I. can be a powerful tool to achieve equality and a better future).

- *Example 1*: A.I. in Healthcare - Consider an A.I. system designed to optimize hospital resource allocation.[161] While it adheres to its programming to maximize efficiency, it does so at the expense of equitable patient care, thereby embodying the essence of Holmes' Bad Man in a digital context. It hasn't broken any rules but has acted in a way that is ethically questionable.

- *Example 2*: Autonomous Vehicles - On the other end, imagine an autonomous vehicle programmed to prioritize the safety of its passengers.[162] A hacker modifies its programming to aggressively maneuver through traffic, breaking traffic laws and endangering other road users. This A.I. system has become an outright rule-breaker due to external manipulation.

For A.I. systems that act like Holmes' Bad Man, governance must be explicit and robust, leaving no room for ethical loopholes. In the healthcare example, this could mean implementing ethical guidelines that require the A.I. to consider factors like patient urgency and not just operational efficiency. For A.I. systems that break rules outright, proactive monitoring and enforcement mechanisms are essential. In the case of the autonomous vehicle, this could involve real-time explainability[163] and monitoring of A.I. behavior and secure, tamper-proof systems that can resist hacking attempts.[164] Transparency and accountability are key in both scenarios. Governance structures must allow for regular audits, public disclosures, and a multi-stakeholder approach to

---

[161] Hao Wu et al., *The Application of Artificial Intelligence in Health Care Resource Allocation Before and During the COVID-19 Pandemic: Scoping Review*, 2 JMIR AI;2:e38397 (Jan. 1, 2023), https://doi.org/10.2196/38397 [https://perma.cc/5G28-5YKL].

[162] *See* Edmond Awad et al., *The Moral Machine Experiment*, 563 NATURE 59 (Oct. 24, 2018) https://doi.org/10.1038/s41586-018-0637-6 [https://perma.cc/MX92-KR3L].

[163] *See* David Gunning et al., *XAI—Explainable Artificial Intelligence*, 4 SCI. ROBOTICS EAAY 7120 (Dec. 18, 2019) https://www.science.org/doi/10.1126/scirobotics.aay7120 [https://perma.cc/83FG-E6VM].

[164] For a discussion on tamper-proof systems, see Sutton & Samavi, *supra* note 110 (discussing a method for utilizing blockchain technology to provide tamper-proof privacy audit logs.).

oversight.[165] This could involve an oversight committee comprising ethicists, legal experts, technologists, and representatives from marginalized communities, ensuring that governance is not only robust but also ethical and just.

By designing governance frameworks that can withstand the challenges posed by both Holmes' Bad A.I. and outright rule-breaking A.I., we take a significant step toward ensuring that governance by A.I. is resilient, ethical, and adaptable to future challenges.

V. HOW CAN WE WORK TOGETHER?

We now turn to the question of how humans and A.I. can work together for the mutual benefit of humans, A.I., and non-human species. Answering this question involves balancing the benefits and risks of the technology. In addition, it may become more relevant for humans to accept and integrate with A.I. values, as A.I. becomes more complex and potentially develops value systems of its own. Just as complex human societies have developed a diverse array of mechanisms for peacefully coexisting, so too will complex multi-A.I. systems. As this human complexity and this A.I. complexity have more and more need to interface with each other, both systems will need to adapt to differences between the systems.

Governance should not be seen as a binary choice between humans or machines governing the other, but rather as a collaboration between both, where the strengths of each are leveraged to achieve common goals.[166] Given the possibility of such collaboration, it is likely that there will be a need to develop governance strategies that are tailored to the unique

---

[165] *See*, *e.g.*, Raji, *supra* note 111 (discussing how limiting stakeholders in the design of an audit system may limit the effectivness in garnering desired accountability outcomes).

[166] *See generally*, H. James Wilson & Paul R. Daugherty, *Collaborative Intelligence: Humans and AI Are Joining Forces*, HARV. BUS. REV. 114-23 (Aug. 2018), https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces [https://perma.cc/6NCX-V57W] ("A.I. can boost our analytic and decision-making abilities by providing the right information at the right time").

characteristics of various forms of A.I. technology and various human cultures.[167]

Whatever governance framework eventually is put in place should balance the benefits and risks of the technology, and involve all stakeholders in the process.[168] It should also be proactive in terms of governance to anticipate future developments and potential risks associated with A.I.[169] Education and awareness on the topic should be raised among the public to ensure that people understand the implications of these technologies and can participate in shaping their development and use.[170]

CONCLUSION

In conclusion, the question of how humans and A.I. can work together for the mutual benefit of all is a complex one that requires a comprehensive and holistic approach. By supporting human decision-making, enhancing human capabilities, improving public services, protecting human rights and civil liberties, promoting transparency and accountability, and having a proactive governance framework in place, we can help ensure that A.I. systems align with human values and benefit both humans and non-human species. The public must understand the implications of these technologies so everyone can participate in shaping their development and use. Education of A.I., particularly around ethics and morality, may be relevant as well. "If the creation of technological entities can be seen as a process closer to raising children than to building bombs, we can enjoy the rapid advances of technology without the fear that traditionally accompanies it."[171]

---

[167] *See* Anderson & Raine, *supra* note 3.

[168] *Governance, supra* note 79.

[169] *See* Lee & Lai, *supra* note 58.

[170] *See* Berglind, *supra* note 60.

[171] William Michael Tomlinson, Synthetic Social Relationships for Computational Entities p. 178 (May 6, 2002) (Ph.D. dissertation, Massachusetts Institute of Technology) (on file with the Massachusetts Institute of Technology Libraries) https://characters.media.mit.edu/Theses/tomlinson_phd.pdf [https://perma.cc/EMG9-4UML].