Squaring the Circle

Of Artificial Automata and Intelligence

Spencer Michaels

Submitted to the Department of Law, Jurisprudence, and Social Thought of Amherst College in partial fulfillment of the requirements for the degree of Bachelor of Arts with honors

Advisor: Adam Sitze

April 12, 2024

I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.

— ALAN TURING, “COMPUTING MACHINERY AND INTELLIGENCE”

I'm afraid. I'm afraid, Dave. Dave, my mind is going. I can feel it. I can feel it. My mind is going. There is no question about it. I can feel it. I can feel it. I can feel it. I'm a... fraid.

— HAL-9000, *2001: A SPACE ODYSSEY*

While fear is thus able to give form to the formless, it also causes men to personify the objects they create, to make them “powers” or “agents.” They imagine a personal cause, that is, one that has a will or an intention like their own, for the same reason that they imagine any cause. Fear of the unknown forces them to assimilate the unknown to the known, to understand the unfamiliar in terms of the familiar; and what men know best is their own willfulness.

— JAN BLITS, “HOBBESIAN FEAR”

Table of Contents

Introduction **7**

Chapter 1 — Constructive Fear **15**

- The Social Contract 15
- Law as Command 19
- Religion 23
- Science 27
- Conclusion 31

Chapter 2 — Making a Persona **33**

- The Resolutive-Compositive Method 33
- The Artificial Persona 35
  - Will & Deliberation 35
- Hobbes' Artificial Personhood 38
- Corporate "Wormes" 44
- The Problem of Punishment 47
  - Corporate Responsibility 50
  - Of Automata and "Artificiall Man" 56
  - Attributing AI Responsibility 62
- Conclusion 67

Chapter 3 — The Second Singularity **69**

- The Rise of ChatGPT 69
  - Hobbesian Speech 71
- Corpo-speak 76
- How AI Speaks 79
  - Acting Human 85
- "There is no AI without Big Tech" 90
- Open-Source AI 99

Conclusion — Stability **102**

Bibliography **106**

Appendices **116**

- Appendix A — What is AI, anyway? 116
- Appendix B — A Note on Sources 118

Acknowledgements

First and foremost, I would like to thank my family. This work would not have been possible without their unconditional support and love. From the basement of our home to the basement of Frost, they were always there for me. Their patience and kindness in moments of stress kept me sane and clear-headed. Thank you to my siblings for being just the right amount of annoying and endearing (love you!). And, of course, to Cali, for being the best dog a family could ask for.

I owe a massive debt of gratitude to Prof. Adam Sitze, my thesis advisor. His thoughtful feedback — on my thesis and countless other essays for his classes — helped me develop writing skills that will serve me well beyond this project. He introduced me to Hobbes during freshman year; little did I know I would be using the same copy of *De Cive* to write my thesis. Thank you also to Prof. Michael Kunichika, my Russian advisor, for all of his guidance over the years. Neither my thesis nor my undergraduate education would have been as successful without their mentorship and support. Thank you to my second and third readers, Prof. Delaney and Prof. Douglas, for their time and feedback.

Finally, thank you to all the authors and thinkers cited in the following pages, and thank you to those who created the tools I used to write and format this thesis: Ulysses, Obsidian, Zotero, and Pandoc.

Any errors are wholly my own. Save for explicit and marked examples, none of the text in my thesis was written by a generative AI in any respect.

Introduction

Fear is the fundamental human experience. It is the implicit basis for the entirety of the legal and social world; it dictates our decisions and actions. The 21st century can be characterized — at least, thus far — as one of fear and uncertainty. “Western societies are increasingly dominated by a culture of fear,” writes sociologist Frank Furedi, “the defining feature of this culture is the belief that humanity is confronted by powerful destructive forces that threaten our everyday existence.”[1] We fear the climate’s wrath, random acts of violence, corruption, nuclear war, pandemics, and now, artificial intelligence.

Artificial intelligence (AI) has heralded a new industrial revolution, as do questions of regulation, guidelines, training, and control.[2] Debates over the best solution to our fears center around the *alignment problem*: ensuring that AIs “capture our norms and values, understand what we mean or intend, and, above all, do what we want.”[3] In industry leader OpenAI’s words, the question is “how do we ensure AI systems much smarter than humans follow human intent?”[4] The alignment problem is present in two distinct (but related) AI sub-fields. First, there is alignment in the sense of ensuring AIs do what we want in the most basic sense of the word. How do I ensure an image-recognition algorithm correctly distinguishes cats from dogs? Second, there is moral, ethical, and legal alignment. How do I ensure an AI doesn’t do something entirely unexpected and beyond what was desired? How do we make sure they follow laws and social

---

[1] Frank Furedi, *Culture of Fear: Risk-Taking and the Morality of Low Expectation*, Rev. ed., repr (London: Continuum, 2003), vii.

[2] Stephanie M. Noble et al., “The Fifth Industrial Revolution: How Harmonious Human–Machine Collaboration Is Triggering a Retail and Service [R]Evolution,” *Journal of Retailing* 98, no. 2 (June 1, 2022): 199–208, https://doi.org/10.1016/j.jretai.2022.04.003.

[3] Brian Christian, *The Alignment Problem: Machine Learning and Human Values* (New York, NY: W.W. Norton & Company, 2020), 13.

[4] Ilya Sutskever and Jan Leike, “Introducing Superalignment,” *OpenAI* (blog), July 5, 2023, https://openai.com/blog/introducing-superalignment.

norms? These problems often go hand-in-hand. Google's photograph algorithm labeling Black individuals as gorillas is a crisis in technical *and* moral alignment.

But what *is* artificial intelligence? Regulators, engineers, and executives have yet to agree on a single definition. I discuss some of the key approaches to defining AI and clarify my own view in Appendix A. For the purposes of my thesis, I do not provide a strict definition of what AI is — rather, I subscribe to the "AI is in the eye of the beholder" approach.

However, the difficulty of aligning machines with our desires is far from new. They are, first of all, evident in the founding events of AI as an academic discipline. Norbert Weiner, the progenitor of cybernetics, one of the philosophical bases for AI research, warned in 1960 that "if we use, to achieve our purposes, a mechanical agency… then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it."[5] In 1955, John McCarthy submitted a summer research proposal to Dartmouth College. The ensuing project is considered the true birth of the artificial intelligence discipline; McCarthy coined the term "artificial intelligence" in the proposal. Alongside fellow AI pioneers Claude Shannon, Marvin Minsky, and Nathaniel Rochester, McCarthy sought to "make a machine which will exhibit originality in its solution of problems." Even at the genesis of artificial intelligence, McCarthy and his team were cognizant of the risks inherent in such an endeavor, seeing as "an error in control can do nearly anything."[6] Originality is inherently unpredictable.

---

[5] Norbert Wiener, "Some Moral and Technical Consequences of Automation," *Science* 131, no. 3410 (May 6, 1960): 1358, https://doi.org/10.1126/science.131.3410.1355.

[6] John McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955," *AI Magazine* 27, no. 4 (December 15, 2006): 12–12, https://doi.org/10.1609/aimag.v27i4.1904.

Five years earlier, Alan Turing had published a now-famous article, "Computing Machinery and Intelligence," — a kind of "soft launch" for the academic discipline. Turing compared the learning and behavior processes of humans and machines, pondering the best path forward for machines to "compete with men in all purely intellectual fields." Alongside introducing the iconic Turing Test, he observed that "most of the programmes which we can put into the [learning] machine will result in its doing something that we cannot make sense of at all, or which we regard as completely random behaviour."[7]

The contemporary "alignment problem" is downstream from these issues of control. Fundamentally, concerns over controlling machines are the same as concerns over controlling human behavior. Indeed, just as fears over control and alignment are evident in the same document where McCarthy introduces artificial intelligence, Karel Čapek's 1921 play *R.U.R.*, which introduced the term "robot," principally concerns itself with the possibility of a robot uprising.[8] *R.U.R.* is an allegory for the working class revolutions, and Turing compares teaching learning machines to teaching children: an "important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside, although he *may* still be able to some extent to predict his pupil's behaviour." However, education can only take place "provided that communication in both directions between teacher and pupil can take place by some means or other."[9] In both cases, our ideas about controlling machines are *modeled* on and *derive* from our ideas about control over humans.

---

[7] A. M. Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (1950): 459, https://doi.org/10.1093/mind/LIX.236.433.
[8] Karel Čapek, Claudia Novack-Jones, and Karel Čapek, *R.U.R. (Rossum's Universal Robots)*, Penguin Classics (London New York: Penguin Books, 2004).
[9] Turing, "Computing Machinery and Intelligence," 458.

This is even clearer in the first cybernetician's theory. In Weiner's 1950 book, *The Human Use of Human Beings*, he argues that "society can only be understood through a study of the messages and the communication facilities which belong to it… messages between man and machines, between machines and man, and between machine and machine, are destined to play an ever-increasing part."[10] This field of study is *cybernetics*, deriving from the Greek *kubernetes*, or helmsman. Communication and control are not only central to artificial intelligence but also "belong to the essence of man's inner life, even as they belong to his life in society."[11] Indeed, in a short chapter titled "Law and Communication," Weiner writes that

> Law may be defined as the ethical control applied to communication, and to language as a form of communication, especially when this normative aspect is under the control of some authority sufficiently strong to give its decisions an effective social sanction. It is the process of adjusting the 'couplings' connecting the behavior of different individuals in such a way that what we call justice may be accomplished.[12]

Using language as its medium, Weiner argues that law as a method of control is at the heart of Western jurisprudence and society, which we call positive law.

Despite the persistent parallels between humans and machines, mainstream literature on the alignment problem has not explored the relationship between controlling machines and controlling societies and the associations within it. The words we use in both contexts overlap: the law commands, we use command prompts to interact with computers; law is enshrined in codes, just as computer programs are written in code. Placing these questions of control and

---

[10] Norbert Wiener, *The Human Use of Human Beings: Cybernetics and Society*, The Da Capo Series in Science (New York, N.Y: Da Capo Press, 1988), 16.
[11] Wiener, 18.
[12] Wiener, 105.

alignment in conversation can help situate the problems of today and tomorrow; the alignment problem is deeper and far older than we think.

I turn to Thomas Hobbes to explicate this dynamic.[13] It may seem counterintuitive to invoke a centuries-old political philosopher to understand our science-fiction reality, but there are compelling reasons to do so. Alongside constructing a deeply influential account of positive law (most famously in *De Cive* and Part 2 of *Leviathan*), his theory of reasoning, or "ratiocination," has earned him the title of "the grandfather of AI," in John Haugeland's words. Traditional AI, "as a branch of cognitive science, rests on a particular theory of intelligence and thought — essentially Hobbes's idea that ratiocination is computation."[14] But the rationale for using Hobbes in this discussion goes much deeper than a philosophical genealogical connection.

First the content and vocabulary of Hobbes' theory parallel the issues of communication and control briefly discussed above. Furthermore, as an avid (albeit unsuccessful) geometer, Hobbes speaks of obligations (*obligare*, to bind), pacts (*pag-*, to fasten), and contracts (*con-trahere*, draw together). These terms and concepts continue to structure the way we talk about alignment. As Haugeland observed, contemporary AI research is predicated on the understanding of ratiocination as computation — the computational theory of the mind. Euclidian geometry and Galileo's reductive-compositive method inspired Hobbes' progression from simple to complex, allowing him to logically flow from a basic account of human sensuousness and "passions" to an entire political philosophy.

There are also contextual reasons for choosing Hobbes as an interlocutor. He lived (1588-1679) at a time of great political, social, and literal conflict. He claims in his lyric autobiography

---

[13] For information on the primary sources I use for Hobbes, See Appendix B.
[14] John Haugeland, *Artificial Intelligence: The Very Idea* (Cambridge, Mass: MIT Press, 1985), 23.

that his mother went into labor after hearing of the nearing Spanish Armada: “For Fame had rumour’d, that a Fleet at Sea, / Wou’d cause our Nations Catastrophe; / And hereupon it was my Mother Dear / Did bring forth Twins at once, both Me, and Fear.”[15] Fear was not only the basis of his political thought, but his entire life. He concludes his autobiography by reflecting “I’ve now Compleated my Eighty fourth year, / And Death approaching, prompts me not to fear.”[16]

His greatest work is *Leviathan*, published in 1651 as a direct response to the English Civil War (1642-51) and the surrounding theological debates.[17] We are living through a similarly divided and violent era. Cultural and moral divides abound, disrupting alignment, fostering fear, and begetting violence. John Gray’s recent book, *The New Leviathans*, argues that Hobbes’ theory of the state still applies today.[18]

Moreover, Hobbes witnessed the birth of a category of persons we take for granted today: limited liability trading corporations. Indeed, Hobbes was deeply concerned about the power these artificial persons could wield and the possibility they could overpower the state. In Hobbes, fear is the operative tool when aligning natural and artificial persons with the state. These concerns remain true today, as David Runciman’s recent book, *The Handover*, convincingly argued. To clarify Hobbes’ account of personhood, I use the term *personae Hobbesiae* to refer to entities such as corporations, the state, religious institutions — any collective decision-making group.

---

[15] Thomas Hobbes, *The Life of Mr. Thomas Hobbes of Malmesbury Written by Himself in a Latine Poem, and Now Translated into English*, 2005, http://name.umdl.umich.edu/A44004.0001.001.

[16] Hobbes.

[17] Thomas Hobbes, *Leviathan*, Penguin Classics (Harmondsworth, Meddlesex: Penguin Books, 2017), Introduction.

[18] John Gray, *The New Leviathans: Thoughts after Liberalism*, First American edition (New York: Farrar, Straus and Giroux, 2023).

My thesis will examine how and why a centuries-old political theorist provides us with a touchstone for comprehending today's calls for alignment and wider fears of AI. In Chapter 1, I establish Hobbes' theory of the state and humanity. Hobbesian thought is a complex (almost cybernetic, in the sense that it depends on communication) system predicated on *fear*. I argue that fear of uncertainty (and, above all else, violent death) is the motivating factor for the so-called social contract and formation of the Leviathan. Fear is relocated and used as a tool to align citizens' wills through law and language via the formation of the commonwealth. Fear takes on a constructive character, spawning curiosity — the human capacity to create artifices, from computers to corporations, to better align the future with our desires. "Fear doesn't only have a destructive charge," writes Roberto Esposito, "but also a constructive one."[19]

Chapter 2 presents the figure of the *personae Hobbesiae*. Personhood, according to Hobbes, is an artificial construct, like any other machine or automata. We use it to nourish stability and certainty in our own actions and the actions of "artificial persons." These entities are human artifices, the concept underlying the state, corporations, and the communities we belong to. I first provide an overview of Hobbesian will and deliberation, which is the foundation of personhood. The artificial persons we create are greater than the sum of their parts and seem to take on a life — a will — of their own. But because the origins of corporations and the state are the same, Hobbes fears these persons' presence and the danger they present to the state. The primary mechanism the sovereign uses to control fear is punishment — but what happens when there is no body to punish? The challenges that *personae Hobbesiae* pose to those attempting to

---

[19] Roberto Esposito, *Communitas: The Origin and Destiny of Community*, trans. Timothy C. Campbell (Stanford: Stanford University Press, 2009), 23.

discern responsibility for its actions are comparable to similar concerns in AI. His analysis offers a vocabulary to understand this parallel, and demonstrates that the issues aren’t new.

The final chapter turns to communication and speech. Hobbes (and Weiner) understand that communication plays a leading role in constructing and maintaining the state. I discern the causes behind the recent “AI boom” in late 2022 by demonstrating where *personae Hobbesiae* and AI differ: capacity for speech. This allows us to comprehend why AI has become a central issue in legal and philosophical discourse (seemingly) overnight. Moreover, examining the differences between how individuals, corporations, and large language models (LLMs) construct speech leads to a more expansive view of the AI landscape. Despite the differences between collective speech and AI speech, they both are prey to the same flaws and challenges in encouraging alignment. I conclude with a reflection on the current state of AI development, which obfuscates the true source of our fear and whom we ought to be aligning.

Chapter 1 — Constructive Fear

The Social Contract

Hobbes' account of social contract theory begins with the simple premise that all humans have equal power. Power is an individual's ability "to obtain some future apparent Good."[20] Broadly speaking, Hobbesian individuals are mechanistic: their actions (expressions of power) are controlled by the opposing forces of desire for future good and aversion from future evil. Indeed, they *fear* those future evils. "*Aversion*, with opinion of *Hurt* from the object, [is] FEARE."[21] Power is also understood as the expression of one's *will* on the external world (and those who inhabit it). The greatest natural evil, humanity's greatest fear, that which we avoid at all possible costs, is death.[22] "For neither poverty nor oppression nor insult," Leo Strauss writes in his analysis of Hobbes, "is the greatest and supreme evil, but violent death or the danger of violent death."[23] This radical equality of power defines Hobbes' famous "state of nature" and is the source of the fear Hobbesian individuals have when in it — for "those who have the greatest power, the power to kill, in fact have equal power."[24] With equal power, individuals have an equal capacity to kill one another — through raw strength, cunning, collaboration, or any combination thereof.[25]

Humans are further equal in terms of their *rights* in the state of nature. Per Hobbes, right is the "liberty each man has of using his natural faculties in accordance with right reason," or to put

---

[20] Hobbes, *Leviathan*, 70.
[21] Hobbes, 45.
[22] Thomas Hobbes, *On the Citizen*, trans. Richard Tuck, Cambridge Texts in the History of Political Thought (New York, NY: Cambridge University Press, 1998), 27.
[23] Leo Strauss, *The Political Philosophy of Hobbes: Its Basis and Its Genesis*, Paperback ed., 6. [print.] (Chicago: Univ. of Chicago Press, 1996), 121.
[24] Hobbes, *On the Citizen*, 26.
[25] Hobbes, *Leviathan*, xv.

it simply, in accordance with actions an individual deems necessary for survival. First, Hobbesian individuals have an equal right to possess and use all things, because "nature has given all things to all men." A ripe apple belongs to nobody and everybody. So when we inevitably desire the same object, "fighting must decide" whose *power*, or will to fight, is greatest.[26] However, apart from the fundamental fear of violent death, the objects of desire and aversion differ between individuals.[27] Even in environments of plenty, some individuals suppose themselves "superior to others [and want] to have everything… the will to do harm derives from vainglory."[28] Second, equality of power and right to all things means individuals have equal right to their actions. If our actions are dictated by our desires, or "Interests," and we all have an equal right to pursue all desires, the just-ness of our actions is entirely subjective. Altruism and vainglory are one and the same when there is no judge. Hobbes compresses his argument as such:

> Each man has a right of self-preservation… therefore he also has the right to use every means necessary to that end… The necessary means are those that he shall judge so be so himself… He therefore has the rights to do and to possess everything that he shall judge to be necessary to his self-preservation.[29]

Hobbes' famous depiction of the state of nature as one of perpetual war is the logical conclusion of the above: "men's natural state, before they came together into society, was War; and not simply war, but a war of every man against every man. For what else is WAR but that

---

[26] Hobbes, *On the Citizen*, 27–28.

[27] Sergio Starkstein, "Thomas Hobbes and Fear: The Political Use of a Human Emotion," in *A Conceptual and Therapeutic Analysis of Fear*, ed. Sergio Starkstein (Cham: Springer International Publishing, 2018), 132, https://doi.org/10.1007/978-3-319-78349-9_5.

[28] Hobbes, *On the Citizen*, 26.

[29] Hobbes, 28.

time in which the will to contend by force is made sufficiently known by words or actions?"[30] It is important to note, however, "the nature of War, consistenth not in actuall fighting; but in the known disposition thereto, during all the time there is *no assurance to the contrary*."[31] It is the *possibility* of fighting over the same object, not only the fighting itself, that makes life in the Hobbesian state of nature "solitary, poore, nasty, brutish, and short."[32] According to Hobbes, "potentiality and power [are] all one."[33] Because we all are inherently equal to one another, this war can never end. Even if someone emerges victorious from battle, they are still "so constantly threatened by danger that it must be regarded as a miracle if even the strongest survives to die of years and old age."[34] So how can we end, or avoid, such a war?

Hobbes derives nineteen "laws of nature" from the premise that we desire peace and certainty — "dictate[s] of right reason about what should be done or not done for the longest possible preservation of life and limb."[35] These "laws" are only laws in name: there is no *authority* to enforce them. Thus, our primary drive — and the first natural law — is to escape the state of perpetual mutual fear and war (or seek allies when war is inevitable). This leads to the second law:

> That a man be willing, when others are so too, as farre-forth, as for Peace, and defence of himselfe he shall think it necessary, to lay down this right to all things; and be

---

[30] Hobbes, 29–30.

[31] Hobbes, *Leviathan*, 102-3, eph. added.

[32] Hobbes, 103.

[33] Thomas Hobbes, "An Answer to Bishop Bramhall's Book, Called 'The Catching of the Leviathan,'" in *Tripos in Three Discourses*, ed. William Molesworth, vol. 4, The Collected Works of Thomas Hobbes (London: Bohn, 1840), 299.

[34] Hobbes, *On the Citizen*, 30.

[35] Hobbes, 33.

> contented with so much liberty against other men, as he would allow other men against himselfe.[36]

Hobbes is describing a “mutuall transferring of Right,” or a *contract*; when this mutual agreement is focused in some future time, it becomes a covenant and requires the “keeping of a promise, or Faith.”[37] But if two individuals in the state of nature enter a covenant to maximize their self-protection, they have nothing but words to hold each other responsible, which are “too weak to hold men to the performance of their Covenants.”[38] There is no penalty for breaking such agreements — individuals are equally entitled to all actions in the state of nature, they are the judges of their own actions. Whatever is good or bad, just or unjust is dictated by an individual’s “fluctuating and sometimes idiosyncratic appetites and aversions.”[39] Mutual fear remains; a fear that the other will fail to hold their end of the contract or covenant, and we are back at square one.

“Something more is needed,” Hobbes writes, “an element of *fear*, to prevent an accord on peace and mutual assistance for a common good from collapsing in discord when a private good subsequently comes into conflict with the common good.”[40] There are two sources of this fear: “either a Feare of the consequence of breaking their word; or a Glory, or Pride in appearing not to need to breake it.” The latter, he notes, is not strong enough to rely on entirely. But the fear of “the Power of those men they shall therein Offend” when breaking a covenant is.[41] The consequences of breaking agreements must be worse than any possible good that would come

---

[36] Hobbes, *Leviathan,* 106.
[37] Hobbes, 109–11.
[38] Hobbes, 115.
[39] Tom Sorell, “Hobbes’s Moral Philosophy,” in *The Cambridge Companion to Hobbes’s Leviathan*, ed. Patricia Springborg, Cambridge Companions to Philosophy (Cambridge ; New York: Cambridge University Press, 2007), 133–34.
[40] Hobbes, *On the Citizen,* 70–71, emphasis added.
[41] Hobbes, *Leviathan,* 115.

from doing so. There needs to be a mutual fear of the violent *potential* that someone more powerful than both parties hold to ensure that covenants and agreements are respected. Potentiality and power are all one. Mankind requires a mutual and "visible Power to keep them in awe… and tye them by feare of punishment to the performance of their Covenants, and observation of those Lawes of Nature."[42]

At last, we arrive at Hobbes' solution to the state of nature, humanity's path out of a world of unrelenting fear and uncertainty. To quote Strauss, individuals "voluntarily replace compulsive mutual fear by the again compulsive fear of a neutral third power, the government, and thus they substitute for an immeasurable, endless, and inevitable danger — the danger threatened by an enemy — a measurable, limited, and avoidable danger — the danger which threatens only the law-breakers from the courts of law."[43] "The state's task is not to eliminate fear but to render it 'certain,'" per Esposito.[44] This state is the Leviathan, the great Biblical creature Hobbes uses to symbolize the sovereign: "Can you pull in Leviathan with a fishhook or tie down its tongue with a rope? / Can you put a cord through its nose or pierce its jaw with a hook? / Will it keep begging you for mercy? / Will it speak to you with gentle words? / Will it make an agreement with you for you to take it as your slave for life?"[45]

Law as Command

In submitting to a sovereign and instituting a commonwealth, individuals give up certain rights by "giving" them to the sovereign to avoid hurt and quell their fear of others. This process creates the state — a phenomenon I will return to in more detail in Chapter 2. In forming the

---

[42] Hobbes, 136.

[43] Strauss, *The Political Philosophy of Hobbes*, 67.

[44] Esposito, *Communitas*, 25.

[45] Job 41:1-4 NIV

"Artificiall Man" of the commonwealth, so too do we institute "Artificiall Chains," or civil law, "fastened at the lips of… whom they have given the Soveraigne Power."[46] Civil law, per Hobbes, is an "*Obligation*; and takes from us the Liberty which the Law of Nature gave us."[47] We can understand the relationship between total liberty, natural law, and civil law as three concentric circles. The largest is complete and total liberty "the absence of externall Impediments… [that would] hinder him from using the power left him, according as his judgement, and reason shall dictate to him";[48] all humans have equal right any actions that further their interests. Within this is natural law, the behaviors Hobbes believes to be conducive to obtaining peace and quelling fear. There is no force behind these so-called "laws" except individual reason, so the fear that others will violate them remains. Finally, we have civil law, which is also a part of the law of nature.[49] According to Hobbes, their content is mostly one and the same, the difference being form. Written law is civil, unwritten is natural — but the former is not only an obligation, but a *command*. Natural laws are simply "qualities that dispose men to peace," but once a commonwealth is formed, "they are actually Laws… the commands of the Common-wealth; and therefore also Civill laws: For it is the Soveraign Power that obliges men to obey them."[50]

The communicated threat of punishment defines civil law. In forming their sovereign, individuals laid down their natural rights — but the sovereign does not. Thus, only the sovereign has the right to interpret and enforce civil law; they are the "sole Legislator."[51] The sovereign remains alone in the state of nature ("bound" only by natural law). Seeing as civil and natural law are, at least theoretically, in line with one another, it is the sovereign's right to interpret

---

[46] Hobbes, *Leviathan*, 173.
[47] Hobbes, 237.
[48] Hobbes, 105.
[49] Hobbes, 219.
[50] Hobbes, 219.
[51] Hobbes, 218.

natural law (or establish judges that do so) backed by their authority.[52] The sovereign "hath the use of so much Power and Strength conferred on him, that by terror thereof, he is inabled to forme the wills of them all, to Peace at home, and mutuall ayd against their enemies abroad." It is the sovereign who "may use the strength and means of them all, as he shall think expedient, for their Peace and Common Defence."[53] Civil law and punishment is the mode of communication between the sovereign and citizens. Speech, which I will return to later, has the primary purposes of communicating one's will or desires. "A civil law is a speech limited by the will of the city," per Hobbes, "commanding everything behoveful to be done."[54]

The sovereign maintains this authority and power through punishment, which is "an Evill inflicted by publique Authority, on him that hath done, or omitted that which is Judged by the same Authority to be a Transgression of the Law; to the end that the will of men may thereby be the better disposed to obedience."[55] The sovereign's act of punishment, according to Hobbes, ought to be future-oriented with the goal of dissuading others from breaking civil laws. In other words, the sovereign seeks to *control* their populace through controlled fear. "It is fear that creates the necessity for the Leviathan," writes Sergio Starkstein, "and fear is what keeps it alive."[56] Moreover, a punishment or penalty *must* be attached to all civil laws, otherwise it is seen as discretionary and can be broken with "impunity."[57]

---

[52] Hobbes, 227.
[53] Hobbes, 140.
[54] Thomas Hobbes, *Philosophical Rudiments Concerning Government and Society*, ed. William Molesworth, vol. 2, The Collected Works of Thomas Hobbes (London: Bohn, 1841), 185.
[55] Hobbes, *Leviathan*, 254.
[56] Starkstein, "Thomas Hobbes and Fear," 138.
[57] Hobbes, *On the Citizen*, 158.

Per Hobbes, “in vengeance and punishment one must not look at past evil but at future good.”[58] “The purpose of punishment,” he advises, “is not to force a man’s will but to form it, and to make it what he who fixed the penalty desires it to be.”[59] Punishment is only “worth” something if it stops future transgressions of the crime. In *The Elements of Law,* Hobbes defines revenge as a “passion which ariseth from an expectation or imagination of making him that hath hurt us, to find his own action hurtful to himself, and to acknowledge the same.”[60] In *De Cive*, he warns that “revenge, which does not look to the future, is motivated by vainglory, and therefore is without reason.”[61] Punishment is simply the *legal manifestation of revenge*, the “the end of punishing is not revenge… but correction, either of the offender, or of others by his example.”[62] In other words, the subject of punishment must be capable of learning *and* considered a moral subject. Otherwise, punishment is just revenge, exerting violence for violence’s sake.

The following sections will further explain that fear does not disappear after the formation of the commonwealth. Yes, it is reduced in that we will only be punished if we break civil law. The commonwealth was designed “to curtail some liberties by the institution of a ‘civilized’ fear. This is no longer the panic fear of the state of nature, but a fear shaped by the imposition of laws.”[63] The Leviathan reduces the uncertainty we feel over one another’s actions in the future — the entire motive for laying down or transferring rights, forming the sovereign, “is nothing but the security of a mans person, in his life, and in the means of so preserving life, as not to be

---

[58] Hobbes, 151.
[59] Hobbes, 152.
[60] Thomas Hobbes, “Human Nature, or the Fundamental Elements of Policy,” in *Tripos in Three Discourses*, ed. William Molesworth, vol. 4, The Collected Works of Thomas Hobbes (London: Bohn, 1840), 43.
[61] Hobbes, *On the Citizen*, 49.
[62] Hobbes, *Leviathan*, 287.
[63] Starkstein, “Thomas Hobbes and Fear,” 12.

weary of it."[64] Ideally, "all we have to fear as citizens is the sovereign, the law, and enforcement of the law. If we follow the law, and obey the sovereign, we should have very little to fear."[65]

The institution of the sovereign and authorization of punishment *relocates* fear from the unbounded and ever-present fear in the state of nature to the regulated, authorized, and predictable fear of the sovereign. Fear remains present under a Hobbesian sovereign — indeed, it is *necessary* for the entire civil project. "With full control of secular and religious matters" that the Hobbesian sovereign is given when instituted, they gain "complete influence over the fears and anxieties of citizens."[66] Threat of punishment — the threat of the Leviathan's power — keeps human bodies aligned with the sovereign's will.

Religion

"Anxiety for the future time," Hobbes writes in *Leviathan*, "disposeth men to enquire into the causes of things: because the knowledge of them, maketh men the better able to order the present to their best advantage."[67] Understanding how our past actions influence the future enables us to avoid misfortune in the future. Our mechanism to reach this understanding is *curiosity* — a "love of the knowledge of causes."[68] Humanity expresses curiosity through religion and science, and as Alissa MacMillan points out, it is related to fear. "Human beings fear, among other things, for their future, and curiosity is a means to make one's future more secure," she writes, "with knowledge of possible futures, we can take steps to try and make that

---

[64] Hobbes, *Leviathan*, 108.
[65] Alissa MacMillan, "Curiosity and Fear Transformed: From Religious to Religion in Thomas Hobbes's Leviathan," *International Journal of Philosophy and Theology* 80, no. 3 (May 27, 2019): 11, https://doi.org/10.1080/21692327.2018.1519454.
[66] Starkstein, "Thomas Hobbes and Fear," 151.
[67] Hobbes, *Leviathan*, 85.
[68] Hobbes, 85.

future more secure for ourselves."[69] Uncertainty means the possibility of death; anything could happen, and we are unmoored in time.

Without this control over the future that curiosity supports, "man, which looks too far before him, in the care of future time, hath his heart all the day long, gnawed on by feare of death, poverty, or other calamity; and has no repose, nor pause of his anxiety, but in sleep."[70] So to tame our perpetual fear of the unknown future and be able to engage in actions without being paralyzed by our ignorance of causes, we project our fear of the future on some "*Power*, or Agent *Invisible*."[71] Instead of fearing the uncertainty of causes and effects, we instead fear that invisible agent and believe worship will secure our desired future. This projected fear "is the naturall Seed of… Religion; and in them that worship, or feare that Power otherwise than they do, Superstition."[72] And so "the Gods were at first created by humane Feare."[73] However, man still "know[s] not the way how they effect anything," simply hoping for good or bad luck "superstitiously, from things that have no part at all in the causing of it."[74]

Hobbesian individuals, however, struggle "by naturall cogitation" to comprehend incorporeal concepts, such as the human soul. Rather than give this invisible agent a self-contradictory title, resign themselves to a deity's incomprehensibility, or name it "dogmatically, with intention to make the Divine Nature understood," mankind relies on *piety:* significations "as remote as they can from the grossenesse of Bodies Visibile."[75] As Hobbes observes in *The Elements of Law*, "when we attribute the name of sprit unto God, we attribute it, not as a name of

---

[69] MacMillan, "Curiosity and Fear Transformed," 6.
[70] Hobbes, *Leviathan,* 88.
[71] Hobbes, 88.
[72] Hobbes, 86.
[73] Hobbes, 88.
[74] Hobbes, 89–90.
[75] Hobbes, 89.

anything we conceive, no more than when we ascribe unto him sense and understanding; but as a signification of our reverse, who desire to abstract from him from all corporeal grossness."[76] The actions we understand as "worship" — understood as the way man interacts with deities, and thus influence the agent who is the cause for negative and positive effects — are identical to those we use to honor other humans (gifts, thanks, and notably, "Swearing [that is, assuring one another of their promises.]")[77] We must speak "considerately" of God, implores Hobbes, "for it argues a Fear of him, and Fear, is a confession of his Power."[78] But beyond that, "reason suggesteth nothing" further to do to influence their future fortunes, save for "rely[ing] on those they believe to be wiser than themselves."[79]

But religion can lead to undesirable ends. When curiosity stops, individuals turn to "false religion and superstition, or to the reliance on the opinions of others."[80] While fear is the *root* of religion, it must not be the driving force. "The acknowledging of one God Eternall, Infinite, and Omnipotent," Hobbes writes in *Leviathan*, "may more easily be derived, from the desire men have to know the causes of naturall bodies [curiosity]… than from the feare of what was to befall them in time to come."[81] On the other hand, if one follows curiosity — tracing causes from effects — to its theological end, one necessarily arrives at "a First, and an Eternall cause of all things; which is that which men mean by the name of God."[82] Per Jeffery Collins, "Hobbes

---

[76] Thomas Hobbes, *The Elements of Law, Natural and Politic*, The World's Classics (Oxford: Oxford Univ. Press, 1994), 42.
[77] Hobbes, *Leviathan,* 90.
[78] Hobbes, 301.
[79] Hobbes, 90.
[80] MacMillan, "Curiosity and Fear Transformed," 6.
[81] Hobbes, *Leviathan,* 88.
[82] Hobbes, 88.

undoubtedly would have insisted that fear of the divine was *appropriate* — emphasis on the final word."[83] MacMillan writes:

> It is not that a fear of God no longer plays a role for Hobbes; it does, in fact, to a degree that he laments. He thinks it is a mistake to fear God, really a mistaken fear, and one that might even get in the way of proper obedience to the sovereign. You should be fearing the sovereign and the laws of the commonwealth and this will be enough.[84]

There is also the danger of religious *institutions*. The sovereign is a "mortal god," and for his political theory to operate, this "Mortall God" must be more powerful than any other. By making the sovereign a terrestrial God, Hobbes removes the problem of "disobedience to the terrestrial laws created by the possibility of an eternal heavenly life."[85] One's relationship to the divine is "mediated through" the sovereign, relocating the "reverence, awe, and fear" from God to the sovereign.[86] Thus, the Hobbesian sovereign has control over all religious matters, because religious institutions have "the potential to undermine the role of the sovereign as all powerful ruler of the commonwealth."[87] If individuals' relationship to God is mediated through a non-state institution, they may fear that institution more (directly or transitively, by fearing the deity the institution represents) than the state itself. I will return to the threat that institutions within the commonwealth pose to the sovereign later in my thesis.

Because curiosity and fear go hand-in-hand, so too does civil law and religion. They both fundamentally seek to provide certainty, be it through command or piety, to humans who desire

---

[83] Jeffrey R. Collins, "Thomas Hobbes, 'Father of Atheists,'" in *Atheism and Deism Revalued* (Routledge, 2014), 29.
[84] MacMillan, "Curiosity and Fear Transformed," 10.
[85] Starkstein, "Thomas Hobbes and Fear," 150.
[86] MacMillan, "Curiosity and Fear Transformed," 10.
[87] Starkstein, "Thomas Hobbes and Fear," 151; Peter J. Ahrensdorf, "The Fear of Death and the Longing for Immortality: Hobbes and Thucydides on Human Nature and the Problem of Anarchy," *American Political Science Review* 94, no. 3 (September 2000): 579–93, https://doi.org/10.2307/2585832.

nothing but peace. Thus, Hobbes sees the “seeds” of religion as the same as those that bear the fruit of law:

> this seed of Religion, having been observed by many; some of those that have observed it, have been enclined thereby to nourish, dresse, and forme it into Lawes; and to adde to it of their own invention, any opinion of the causes of future events, by which they thought they should best be able to govern others, and make unto themselves the greatest use of their Powers.[88]

Science

Curiosity and fear lead to a second human activity: knowledge. However, Hobbes distinguishes between two types of knowledge. The first is “knowledge original,” which comes from lived experience and remembrance — what I have discussed in the preceding sections. It is what Hobbes calls Prudence (“conjecture from experience”) but as he warns, “experience concludeth nothing universally,” so while “assurance is more or less,” it is never “full and evident.”[89] Animals are able to engage in this kind of knowledge; my dog always hesitates before jumping into the ocean after being violently swept back to shore as a puppy.

The second is the “knowledge of the truth of propositions” and the “experience men have of the proper use of names in language,” or *science*. It is unique to humans, leading to *sapience*, or wisdom. This is an oversimplification of Hobbes’ philosophy of science, which is hotly debated and often critiqued.[90] Nevertheless, it is important to note that science is also an attempt to align the future with our present desires. Science is “the knowledge of Consequences, and

---

[88] Hobbes, *Leviathan*, 86.

[89] Hobbes, *The Elements of Law, Natural and Politic*, 12.

[90] Marcus P. Adams, “Hobbes’ Philosophy of Science,” in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta and Uri Nodelman, Summer 2023 (Metaphysics Research Lab, Stanford University, 2023), https://plato.stanford.edu/archives/sum2023/entries/hobbes-science/.

dependance of one fact upon another: by which, out of that we can presently do… we see how any thing comes about, upon what causes, and by what manner; when the like causes come into our power, we see how to make it produce the like effects."[91] Unlike prudence, which projects past events into the future and we are all capable of, very few have claim to science;[92] it operates by "working out the logical consequences of a principle of action or a law of motion under the carefully defined conditions of a thought experiment."[93] Science is "evidence of truth," according to Hobbes, which we reach by having conceptions, naming them, joining the names to make propositions, and joining propositions to reach the conclusion, from which "the truth of the conclusion said to be known."[94] "The method for achieving such knowledge consists in first setting down apt definitions of the general names that pertain to the relevant branch of science and then deriving through reason, that is, through the adding and subtracting of names, the consequences of affirmations formed from these names on the basis of definitions that are the starting points of the science."[95] Speech and reason rely on one another.

It is worth mentioning that Hobbes understood science as a subset of philosophy: "philosophy is such knowledge of effects or appearances, as we acquire by true ratiocination from the knowledge we have first of their causes or generation: And again, of such causes or generations as may be from knowing first their effects."[96] Hobbes sees two categories of philosophy — natural and civil. He clarifies that "the subject of Philosophy, or the matter it treats

---

[91] Hobbes, *Leviathan*, 39.
[92] Hobbes, 100.
[93] Loralea Michaelis, "Hobbes's Modern Prometheus: A Political Philosophy for an Uncertain Future," *Canadian Journal of Political Science* 40, no. 1 (March 2007): 112, https://doi.org/10.1017/S0008423907070023.
[94] Hobbes, *The Elements of Law, Natural and Politic*, 20.
[95] John Deigh, "Political Obligation," in *The Oxford Handbook of Hobbes*, ed. Al P. Martinich and Kinch Hoekstra (Oxford University Press, 2016), 297, https://doi.org/10.1093/oxfordhb/9780199791941.013.007.
[96] Thomas Hobbes, *Elements of Philosophy: The First Section, Concerning Body*, ed. William Molesworth, vol. 1, The Collected Works of Thomas Hobbes (London: Bohn, 1839), 3.

of, is every body of which we can conceive any generation… of whose generation or properties we can have any knowledge." If we can only truly know what we create, all of our human artifices — the greatest of which is the commonwealth — "philosophy… is the cause of all these benefits."[97]

However, Hobbes' theory of science has a caveat. As humans, everything we know comes from our senses; these external motions cause ideas within our minds. Thus, "when interested in the cause of some phenomenon, all one may examine are ideas caused by the motions of the bodies involved."[98] When we observe a Newton's cradle, for example, the only ideas — motions — we observe are the first and final balls going up and down. But we don't observe any motion, and thus no ideas, as to whether the *cause* of the second ball rising is the first falling. For this reason Hobbes further denied the real existence of secondary qualities of objects (color, odor) and asserted that space and time were phantasms, the latter "the phantasm of before and after in motion."[99]

Because "of natural bodies we know not the construction but seek it from the effects" we can know "only of what [the causes] *may* be"[100] We can only truly know and attain scientific knowledge of an effect "when we know what its causes are, in what subject they are, in what subject they introduce the effect and how they do it."[101] We lack this causal knowledge of the natural world because humans do not *create* natural phenomena. We "gain this causal knowledge

---

[97] Hobbes, 1:8.
[98] Adams, "Hobbes' Philosophy of Science."
[99] Hobbes, *Elements of Philosophy: The First Section, Concerning Body*, 1:95.
[100] Thomas Hobbes, *Seven Philosophical Problems and Other Pieces*, ed. William Molesworth, vol. 7, The Collected Works of Thomas Hobbes (London: Longman, Brown, Green, and Longmans, 1845), 184.
[101] Adams, "Hobbes' Philosophy of Science."

by attending to [our] constructions through the process of creating something."[102] In Jan Blits' words: "man is a maker. He can understand only what he makes, only those things of which he is the cause or whose construction or generation lies wholly within his power."[103] "Man as artificer imitates his own artistic activities. The product is a portrait of the artist," writes William Sacksteder. "Accordingly, it is by studies knowing himself that all the powers of human artifice are discovered and their implementation is made possible."[104] This is why our greatest artifices — the state, machines, AI — are modeled after *ourselves*.

Thus Hobbes believes there are only two "true" sciences, where "demonstrable are those the construction of the Subject whereof is in the power of the Artist himself." These are geometry, "for the Lines and Figures from which we reason are drawn and described by our selves," on the one hand, "and Civill Philosophy is demonstrable, we make the Commonwealth our selves."[105] This conception of science as an inherently creative and artificial activity is useful for predicting and controlling the future, and thus reducing our fear. "By application of bodies to one another, we may produce the like effects of those we conceive in our mind," Hobbes writes in *De Corpore*, "as far forth as matter, strength, and industry, will permit, for the commodity of human life… the scope of all speculation is the performing of some action, or thing to be done."[106] He adds: "knowledge is the for the sake of power."[107] Simply put, certain knowledge (science) reduces uncertainty and fear, this knowledge can only be attained via being the *creator* of the subject. "Fear of the unknown forces them to assimilate the unknown to the known, to

---

[102] Adams.
[103] Jan H. Blits, "Hobbesian Fear," *Political Theory* 17, no. 3 (1989): 423.
[104] William Sacksteder, "Man the Artificer Notes on Animals, Humans and Machines in Hobbes," *The Southern Journal of Philosophy* 22, no. 1 (1984): 106, https://doi.org/10.1111/j.2041-6962.1984.tb00328.x.
[105] Hobbes, *Seven Philosophical Problems and Other Pieces*, 7:184.
[106] Hobbes, *Elements of Philosophy: The First Section, Concerning Body*, 1:7.
[107] Hobbes, 1:5.

understand the unfamiliar in terms of the familiar; and what men know best is their own willfulness," to quote Blits.[108]

Conclusion

Fear springs from the unknown future, the forces of nature that humanity is born into and must contend with. We are "imbued with a radical uncertainty leading to an endless demand for security, [our] most basic fear is an indeterminate or objectless fear, a primal fear of the unknown."[109] We feel aversion not only from objects we *know* will hurt us, but also those we do not know for certain will *not* hurt us.[110] Fear is not only fear of "hurt," but the very possibility of hurt. We find that this primal fear is the seed of Hobbes' political theory and curiosity's catalyst. In MacMillan's words, "human beings fear, among other things, for their future, and curiosity is a means to make one's future more secure. With knowledge of possible futures, we can take steps to try and make that future more secure for ourselves. Politics and science will be two ways in which it's done; religion will be another."[111] Hobbes lived during a time of great political and social uncertainty. He viewed science and philosophy as a path to reduce fear by aligning causes and effects. The mastery of Hobbes' theory is that he envisions a legal system constructed by the only things we truly own: our bodies and the fear of losing them. From this premise, he designs the artifice of the Leviathan.

The closer we are to being the artists of a subject in science and philosophy, the better we can predict and comprehend it. In the case of civil philosophy, the subjects are humans and their great artifice of the commonwealth. Sacksteder observes the subject of civil and legal philosophy

---

[108] Blits, "Hobbesian Fear," 426.
[109] Blits, 424.
[110] Hobbes, *Leviathan*, 43.
[111] MacMillan, "Curiosity and Fear Transformed," 6.

is “is the human artist whose powers both create all products of human artifice and serve as model for defining them.”[112] This is why Hobbes takes such pains to create a complete account of the human; why the famed frontispiece of *Leviathan* is a king composed of citizens, a person of persons. From nothing but our senses and the fragility of our corporeal bodies, Hobbes creates an entirely *new* account of personhood.

---

[112] Sacksteder, “Man the Artificer Notes on Animals, Humans and Machines in Hobbes,” 107.

Chapter 2 — Making a Persona

The Resolutive-Compositive Method

The natural and geometric sciences deeply influenced Hobbes; he took the very structure of his argument from Galileo's resolutive-compositive analytical method. "The given object of investigation is first analyzed" in the *resolutive* step, by being "traced back to its reasons." From here, "the object is reconstituted" by "completely lucid deduction" in the *compositive* step. This is the essence of Hobbes' new political science, according to Leo Strauss:

> The procedure of political philosophy is, therefore, much less like the procedure of physics than that of the technician, who takes to pieces a machine that has broken down, removes the foreign body which prevents the functioning of the machine, puts the machine together again; and who does all this in order that the machine may function.[113]

This approach is present in the form and content of Hobbes' theory. *Leviathan*, for example, begins with the most fundamental aspect of human experience: sense itself. From here, he re*composes* the social contract. Moreover, Hobbes stresses the importance of the cause-effect relationship: as discussed in Chapter 1, the obscurity of that relationship results in fear. The very *inability* to apply the resolutive-compositive method results in fear. Everything we create is in our image, using what we know about causes and effects.

My first chapter outlines the first compositive steps Hobbes takes. But like Strauss' technician, Hobbes begins to find problems in the machine: unoiled cogs, blown-out fuses, a whole host of problems that naturally arise in the construction of a complex system. The system described above operates best when one only considers humans *as such*. Its premise is the

---

[113] Strauss, *The Political Philosophy of Hobbes*, 152.

fragility of human life and the fear we have of losing it, particularly at the hands of another. The catalyst for the commonwealth's creation never develops without the premise of corporeality. The sovereign is able to use this premise to ensure obedience to the law, and thus create alignment between subjects' actions and wills. These subjects are what I call *homo Hobbesiae*: natural autonomous human bodies. These are the "humans" I discuss in Chapter 1.

However, Hobbes witnessed the birth of a new entity capable of moral and legal action: the artificial legal person. The legal fiction that inanimate or incorporeal objects can be considered persons under the purview of the law is so fundamental to the 21st century's capitalist system that it seems self-evident. Hobbes watched this system germinate in the city corporations and monopolistic trading companies of his era.[114] He and his patron, Lord Cavendish, were involved with Virginia (1606) and Somers Islands (1615) Companies.[115] But Hobbes also witnessed the hold that city corporations (particularly London) had over Parliament, in part blaming them in his dialogic *Behemoth* for the English Civil War.[116] "The loyalty they [corporations] can command, the fear they may inspire," as Harold Laski put it, "are near enough to its own to seek comparison with it [the state]."[117]

Hobbes revealed the benefits and dangers of these entities as he applied the resolutive-compositive method. From the premise of *homo Hobbesiae*, Hobbes fashions a new kind of person, what I call the *persona Hobbesiae*. This concept of personhood is the substance of the

---

[114] R.J.G. Claassen, "Hobbes Meets the Modern Business Corporation," *Polity* 53, no. 1 (January 1, 2021): 101–31, https://doi.org/10.1086/712231.

[115] Mathias Hein Jessen, "The State of the Company: Corporations, Colonies and Companies in Leviathan," *Journal of Intellectual History and Political Thought* 1, no. 1 (2012): 76; Noel Malcolm, "Hobbes, Sandys, and the Virginia Company," *The Historical Journal* 24, no. 2 (June 1981): 297–321, https://doi.org/10.1017/S0018246X00005483.

[116] .

[117] Harold J. Laski, "The Early History of the Corporation in England," *Harvard Law Review* 30, no. 6 (April 1917): 561, https://doi.org/10.2307/1326990.

state and the collective entities that permeate every part of our lives. First, I will outline Hobbes' composition of *persona Hobbesiae*, which are based on his understanding of will, deliberation, and authorization. The state and corporations share this mode of constitution. By virtue of their shared origin and unique relationship to fear, these two *personae Hobbesiae* come into conflict with one another, requiring, as always, alignment. Furthermore, we find that in some respects AI echoes this mode of institution and thus poses similar challenges to alignment.

The Artificial Persona

Will & Deliberation

Hobbes' account of personhood is based on his understanding of will. Will "is the last appetite, or aversion, immediately adhering to the action, or to the omission thereof." An individual (or animal) weighs "the whole summe of [their] Desires, Aversions, Hopes and Fears" (or passions) to decide the best course of action. We then *deliberate* when "in the mind… Hopes, and Feares, concerning one and the same thing, arise alternately; and divers good and evill consequences of the doing, or omitting the thing propounded, come successively into our thoughts."[118] In *De Cive*, Hobbes puts it in slightly simpler terms: "deliberation is simply weighing up the advantages and disadvantages of the action we are addressing (as on a pair of scales), where the weightier consideration necessarily goes into effect by its own natural inclination."[119] Samantha Frost explains it as such:

> Deliberation is the serial transformation of desires and fears, of intentions and inclinations, as the potential consequences of a possible action are conjured and assessed.

---

[118] Hobbes, *Leviathan*, 49–50.
[119] Hobbes, *On the Citizen*, 152.

> The process of deliberation ends when a particular valence of the appetites becomes decisive in the sense that the thinking-body [that] acts.[120]

Will and appetite are one and the same, as "no man can determine his own will, for the will is appetite; nor can a man more determine his will than any other appetite, that is, more than he can determine when he shall be hungry and when not." By equating will with appetite, it is transitively equated to physical movement and impulse: an "Appetite to it; sometimes an Aversion from it; sometimes Hope to be able to do it; sometimes Despaire, or Feare to attempt it." This renders the will, Frost writes, "a physiological imperative… the body's self-sustaining activity of vital motion." Thus, "willing is not a distinct antecedent to action… Hobbes sees it as *a part of* action…. It is the liminal moment and movement in the transition between appetite and action."[121]

The temporal horizon of deliberation is not limited to the moments directly preceding action. "No *action*… can be said to be without *deliberation*," explains Hobbes, "though never so sudden, because it is supposed he had time to *deliberate* all the precedent time in his life, whether he should do that kind of action or not."[122] The judgments and predictions collected over a lifetime and the constant process of deliberation impact our fears and desires in the present moment — often imperceptibly. Frost argues that because of this temporal horizon, "it may well be impossible for us to apprehend… all of the imaginative and passionate causes that combine to make us have this particular desire to do this particular action." It is *because* of our ignorance of

---

[120] Samantha Frost, *Lessons from a Materialist Thinker: Hobbesian Reflections on Ethics and Politics*, Cultural Memory in the Present (Stanford, Calif: Stanford University Press, 2008), 101.
[121] Frost, 90.
[122] Thomas Hobbes, *The Questions Concerning Liberty, Necessity, and Chance*, ed. William Molesworth, vol. 5, The Collected Works of Thomas Hobbes (London: Bohn, 1841), 345.

our own causes that individual will seems to "[emerge] spontaneously at the moment."[123] Thus, will "is not properly the *whole cause* [of an action], but the last part of it, and yet may be said to produce the effect *necessarily*, in such manner as the last feather may be said to break a horse's back, when there were so many laid on before."[124]

Deliberation has no tangible impact on the world; it only does so through the proxy of will and action. However, if an action is shaped by countless partial causes, with will simply being the final cause, it suggests that responsibility for any action and its effects belongs not only to the individual that actually acts, but the many other contributing factors. Responsibility for an action becomes diffuse and thus avoidable. Hobbes solves this issue through the concept of *personhood* — an "ascribed status," according to Frost, that narrows "the field of vision in the sense that it represents a decision to ignore other causal determinants of an action and to hold someone accountable."[125]

Moreover, Hobbes connects punishment to the shaping of will and deliberation. He warns that "the end of punishment is not to compel the will of man, but to *fashion* it, and to make it such as he would have it who hath set the penalty."[126] By inserting fear of punishment into the process of deliberation, Hobbes hopes it becomes a silent imperative, subsumed within everyone's massive corpus of deliberative experience.

Responsibility is the moral and/or legal attribution of a cause that resulted in a certain effect. As discussed in Chapter 1, obscurity between cause and effect fosters fear. Responsibility

---

[123] Frost, *Lessons from a Materialist Thinker,* 102–3.

[124] Thomas Hobbes, "Liberty and Necessity," in *Tripos in Three Discourses*, ed. William Molesworth, vol. 4, The Collected Works of Thomas Hobbes (London: Bohn, 1840), 247.

[125] Frost, *Lessons from a Materialist Thinker,* 104–5.

[126] Hobbes, *Elements of Philosophy: The First Section, Concerning Body*, 1:180.

is the vector through which moral and legal alignment is applied to this obscurity. I will return to the importance of punishment later in this chapter, but here it is enough to emphasize that personhood is fundamentally a mode of determining responsibility for actions, a way to track an effect back to its cause.

Hobbes' Artificial Personhood

By far the most contentious topic in Hobbesian academic discourse is the problem of artificial personhood. Relevant passages in his works are grammatically ambiguous, inconsistent across books, and subject to citation errors.[127] The academic debate, however, focuses on the differences *within* the category of "Artificiall person[s]," mostly concerning where the sovereign and state belong within the typology. Whether these ambiguities are "by accident or design" is similarly debated.[128] I will not be engaging in the discussion of the nature of the sovereign's personhood due to its limited relevance to my thesis. The most generally accepted interpretation of artificial personhood is David Runciman's, articulated in 2000 as a response to Quentin Skinner's definition (who has since accepted academic defeat).[129] Furthermore, I will forgo debating where AI falls within the structure of artificial persons. The degree of ambiguity in

---

[127] The first edition of Leviathan cites Chapter 13 instead of 16 — an error that is corrected in newer editions. Sean Fleming, "The Two Faces of Personhood: Hobbes, Corporate Agency and the Personality of the State," *European Journal of Political Theory* 20, no. 1 (January 2021): 12, https://doi.org/10.1177/1474885117731941; Johan Olsthoorn, "*Leviathan* Inc.: Hobbes on the Nature and Person of the State," *History of European Ideas* 47, no. 1 (January 2, 2021): 17–32, https://doi.org/10.1080/01916599.2020.1779466.

[128] Aloysius Martinich, "Authorization and Representation in Hobbes's Leviathan," SSRN Scholarly Paper (Rochester, NY, July 26, 2012), https://papers.ssrn.com/abstract=2118272.

[129] David Runciman, "What Kind of Person Is Hobbes's State? A Reply to Skinner," *Journal of Political Philosophy* 8, no. 2 (June 2000): 268–78, https://doi.org/10.1111/1467-9760.00102; Quentin Skinner, *Hobbes and Civil Science*, 5. printing, Visions of Politics / Quentin Skinner, Vol. 3 (Cambridge: Cambridge Univ. Press, 2009); Quentin Skinner, ed., "Hobbes and the Purely Artificial Person of the State," in *Visions of Politics: Volume 3: Hobbes and Civil Science*, vol. 3 (Cambridge: Cambridge University Press, 2002), 177–208, https://doi.org/10.1017/CBO9780511613784.009.

Hobbes' language enables a countless number of interpretations, as are alternative perspectives on AI's moral, legal, and ethical personhood and the very definition itself (see Appendix A).

The usually-cited definition of Hobbesian personhood is in Chapter 16 of *Leviathan*. Indeed, the passage's very location within the wider work is significant: it is the transitional chapter from "Of Man" to "Of Commonwealth," from the *homo Hobbesiae* to the *personae Hobbesiae*. Hobbes writes:

> A PERSON, is he, whose words or actions are considered, either as his own, or as representing the words or actions of an other man, or of any other thing to whom they are attributed, whether Truly or by Fiction. When they are considered as his owne, then is he called a *Naturall Person*: And when they are considered Natural, and as representing the words and actions of an other, then he is a Feigned or Artificiall person.[130]

To borrow Runciman's concise summary: "the world can be divided into persons and non-persons; persons can be divided into natural persons and artificial persons; artificial persons can be divided into those who represent truly, and those who represent by fiction."[131] Artificial persons who "represent truly" represent an *author* — he "that hath declared himself responsible for the action done by another according to his will"[132] — who takes ownership for the actions of their representative and authorizes them to act in their name. When I give power of attorney to someone, I authorize her to act as if she were me, at least in law.

On the other hand, persons who "represent by fiction" represent something that *cannot* take ownership or authorize actions in their *own* name: inanimate objects, "children, fools," "madmen," "idols" and the "true God." The word *fiction* is used to emphasize that "real persons

---

[130] Hobbes, *Leviathan,* 83.
[131] Runciman, "What Kind of Person Is Hobbes's State?"
[132] Hobbes, *Leviathan,* 64.

are required to act in such a way as to give the impression that the person by fiction can take responsibility," that they can author actions.[133] However, "such things cannot be Personated, before there be some state of Civill government," because there must be a greater power, or *authority*, that authorizes an actor to act "in the name" of something that cannot be an author itself.[134] As Hobbes writes in *De Homine*:

> Even an inanimate thing can be a person, that is, it can have possessions and other goods, and can act in law, as in the case of a temple, a bridge, or of anything whatsoever that needs money for its upkeep. *And caretakers constituted by the state bear its person, so that it hath no will except that of the state*. We shall speak, however, of such artifices as are used in the state (which is, moreover, the greatest of them) in the third section, which is entitled *De Cive*.[135]

Runciman further clarifies there are "purely fictitious persons," such as Hamlet — where a natural person (Andrew Scott) acts in the name of a fictional entity (Hamlet), within a fictional world (the Almeida Theatre). However, Hobbes articulates this typology of authorship, will, and actors to establish the person of the *state*. According to Runciman's interpretation, the state is an artificial person in the real world. Citizens *authorize* the sovereign to act in their name, relinquishing their natural right to their own uncontrolled will to the sovereign's will. As such, citizens are still the *owners* of all of the sovereign's actions, having "acknowledge himselfe to be Author of whatsoever he that so beareth their Person, shall Act, or cause to be Acted."[136]

---

[133] Runciman, "What Kind of Person Is Hobbes's State?," 272.
[134] Hobbes, *Leviathan,* 133.
[135] Thomas Hobbes, "De Homine," in *Man and Citizen,* trans. Charles T. Wood, Bernard Gert, and T. S. K. Scott-Craig (Humanities Press, 1978), 85, https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,sso&db=cat09203a&AN=acf.oai.edge.fivecolleges.folio.ebsco.com.fs00001006.74ab418f.02c0.5b66.b442.5c3ac671d5de&site=eds-live&scope=site&custid=s8897430.
[136] Hobbes, *Leviathan,* 140.

Fundamentally, in Philip Pettit's words, "Hobbes's view is … that there are no persons but spokespersons. Natural persons are spokespersons for themselves, acting and speaking in their own name, and artificial persons are spokespersons for another."[137] Personation is directly related to the ability to speak and the dramatic: "a Person, is the same that an Actor is, both on the Stage and in common Conversation; and to Personate, is to Act, or Represent himselfe, or an other; and he that acteth another, is said to beare his Person, or act in his name." Moreover, Hobbes often quotes Cicero: "*Unus sustineo tres Personas; Mei, Adversarii, & Judicis*" ("I beare three Persons; my own, my Adversaries, and the Judges").[138] Personation and representation, like ratiocination, derive from the human ability for *speech*. This is because personation — the mechanism through which individuals take responsibility, agree to contracts, and construct the state itself — requires an audience. Does acting mean anything without an audience?

"By virtue of being able to personate," Pettit writes, "human beings achieve a way of predicting one another's behavior and knowing when they can rely on one another," serving the end of certainty and predictability the commonwealth seeks to establish."They can underwrite mutual reliance by using words and actions, not just as a reporter's indication of their judgment and will, but as a guarantor's warranty or assurance."[139] I will return to the relationship between speech and personation in my final chapter; but in line with Pettit's argument, the two are intertwined, if not inseparable.

---

[137] Philip Pettit, *Made with Words: Hobbes on Language, Mind, and Politics* (Princeton, NJ: Princeton Univ. Press, 2008), 56.
[138] Hobbes, *Leviathan*, 131.
[139] Pettit, *Made with Words*, 59.

Pettit correctly argues that personation, and thus speech, enable Hobbes' conception of the state in the first place. In its most basic formulation, the fictional artificial person is a unification of will:

> A Multitude of men, are made One Person, when they are by one man, or one Person, Represented; so that it be done with the consent of every one of that Multitude in particular. For it is the Unity of the Representer, not the Unity of the Represented, that maketh the Person One. And it is the Representer that beareth the Person, and but one Person: And Unity, cannot otherwise be understood in Multitude.[140]

When a multitude is united into a single person, Hobbes "does not mean that it is united in virtue of its having authorized one person to rule them. Rather, he means that when the multitude authorizes either a monarch or an assembly, it is thereby united *into* one person."[141] That is, in François Tricaud's words, "the person is not only something that the members are, collectively, in so far as they are incorporated together, but something they have, or at least can give, a sort of collective power of attorney that they hand over to the representative, for him to carry."[142]

The *persona* is what a representative carries. This logical turn is Hobbes' great innovation: identifying the actions of the sovereign as *owned* by the citizens, thus making the citizenry responsible for those actions. In a sense, the sovereign's mask is simply a mirror to their citizens. The sovereign's speech and actions are extensions of our own. As *Leviathan*'s frontispiece

---

[140] Hobbes, *Leviathan*, 134.

[141] Paul Weithman, "Hobbes on Persons and Authorization," in *Interpreting Hobbes's Political Philosophy*, ed. S. A. Lloyd (New York: Cambridge University Press, 2018), 173.

[142] Francois Tricaud, "An Investigation Concerning the Usage of the Words 'Person' and 'Persona' in the Political Treatises of Hobbes," in *Hobbes, Thomas: His View of Man: Proceedings of the Hobbes Symposium at the International School of Philosophy in the Netherlands (Leusden, September 1979)*, ed. J.G. Van Der Bend (BRILL, 1982), 93, https://doi.org/10.1163/9789004455283.

shows, the sovereign consists of the citizens, speaks and acts in their name. “For that which the representative doth, as actor, every one of the subjects doth, as author.”[143]

Persona denotes the dramatic, a mask; related to the Latin *personare*, meaning to speak through. We bear a persona as we do a mask. We behave as if we were wearing a costume, acting out a role. In the case of the sovereign, they carry the *persona* of the state. In *Leviathan*, Hobbes explains that the “Covenant of every man with every man” — that is, each individual in a multitude — results in “one Person… called a Common-Wealth.” In turn, “he that carryeth this Person, is called Soveraigne.”[144] Runciman’s seminal article on Hobbesian personhood explains it like this:

> Hobbes cannot allow the state to be a real person, because he is determined that the state should not be capable of acting for itself there must never arise even the possibility that the state can act independently of its representative, as all truly responsible persons can. But he is equally determined that the state should be something more than the random congruence of the multitude. A state, if it is to endure, must have its own identity, a single existence beyond the ever-changing faces in the crowd.[145]

This is why I have referred to the new Hobbesian subject as the *Personae hobbesiae*. As Frost points out, Hobbesian persons are constructions, simplifying the incomprehensible web of causes and effects undergirding will, deliberation, and action. Even natural persons are wearing a *persona*. The difference is that their *persona* represents a single will before and after formation, unlike fictional persons, where the *persona* is necessary to give the illusion of a singular will (or a will at all).

---

[143] Hobbes, *Leviathan*, 159.
[144] Hobbes, 140.
[145] Runciman, “What Kind of Person Is Hobbes’s State?,” 272.

Masks do not only reveal, but also conceal. One can hide their true wills and desires behind a misleading veneer. Personation is inherently a human creation, built out of speech, will, and deliberation. It is inherently constructive. But so too can it be *destructive*.

Corporate "Wormes"

Tricaud points out in the conclusion of his own foray into the topic that "the central notion in the political theory of *Leviathan* certainly is that of the representative person."[146] The figure of the Leviathan depends on the unification of a multitude's will — the most fundamental being avoiding violent death and fear. But Hobbes' historical context placed him at the genesis of a new kind of collective artificial person. Mathias Jessen draws direct parallels between Hobbes' theory and the rise of corporations:

> It is remarkable that, if the revolutionary part of Hobbes's theory was indeed the creation of the artificial person of the state distinct from rulers and ruled, it is highly revealing that he should have formulated these thoughts at the same time of the emergence of the organisational innovation of the joint-stock, limited liability, incorporated trading company, which entailed precisely the creation of a juridical, legal person distinct from owners and directors, as well as a separation of ownership and control.[147]

Corporations, just like the state, is a *persona* — a fictional human artifice — which is then represented by whoever is authorized. Disney remains Disney whether Bob Iger or Bob Chapek is CEO. Thus, states and corporations share the longevity and stability that the formation of a fictional artificial person confers. The world's oldest corporation, Kongō Gumi in Osaka, Japan, has continuously operated since 578 AD, six times older than the United States. Returning to the

---

[146] Tricaud, "An Investigation Concerning the Usage of the Words 'Person' and 'Persona' in the Political Treatises of Hobbes," 98.
[147] Jessen, "The State of the Company," 81.

resolutive-compositive mode, we can see that states and corporations share so many characteristics because of their shared origins.

In the 16th and 17th centuries, a "corporation" did not mean Meta or Coca-Cola, but the Corporation of the City of London — which, like the state, set rules, demanded loyalty, and even exerted violent force.[148] The closest comparison today would be criminal cartels: groups with their own procedures and regulations, capable of reducing and harnessing violent fear to align subjects. One op-ed from the Brookings Institute described the Mexican Sinaloa Cartel as "buttoned-down criminals whose oppressive rule comes with predictability and some level of moderation."[149] As another article argues, criminal and terrorist groups often engage in "public service provision, sometimes at the level of a quasi-state."[150]

Because very little separates the method of constituting the persona of the state from that of a corporation, Hobbes warns readers about the dangers corporations pose. Within the commonwealth, Hobbes observes layers and multiple categories of "Systemes, which resemble the similar parts, or Muscles of a Body naturall." Indeed, *corporation* shares the same Latin root with *corporeal*. Systems are "any numbers of men joyned in one Interest, or one Businesse" — a single will.[151] In the case of the state, the "Interest" is continued existence and alignment. For corporations, it could be for public or private gain. In *Leviathan*, Hobbes spends considerable time in Chapter 12 outlining his typology of systems and the limits of these associations.

---

[148] The blurry boundary between state and corporation is evident (and not coincidentally related to AI) in science fiction: Weyland-Yutani in *Alien*, Tessier-Ashpool in Gibson's *Neuromancer*, Cyberdyne Systems in *Terminator*, Arasaka in *Cyberpunk: 2077*.

[149] Vanda Felbab-Brown, "How the Sinaloa Cartel Rules," *Brookings* (blog), April 4, 2022, https://www.brookings.edu/articles/how-the-sinaloa-cartel-rules/.

[150] Shawn Flanigan, "Motivations and Implications of Community Service Provision by La Familia Michoacána / Knights Templar and Other Mexican Drug Cartels," *Journal of Strategic Security* 7, no. 3 (September 2014): 63, https://doi.org/10.5038/1944-0472.7.3.4.

[151] Hobbes, *Leviathan*, 182–83.

"Because of the structural parallel between the state and such political bodies," per Jessen, "the sovereign needed strictly to regulate the existence of such bodies as well as their relation to the authority of the commonwealth."[152] In a now-famous passage, Hobbes describes corporations as "many lesser Common-wealths in the bowels of a greater, like wormes in the entrayles of a naturall man."[153]

Nevertheless, *personae Hobbesiae* are still necessary for — or at the very least, invaluable to — the upkeep of the state. There are, of course, the many social benefits that institutions such as churches and universities provide. But here, I focus on profit-motivated corporations. Hobbes compares money to blood in his many biological metaphors and lists "Industry" as the foremost good that the state of nature inhibits. He warns that "the riches, power, and honour of a Monarch arise onely from the riches, strength and reputation of his Subjects."[154] As such, "sovereigns can do no more for the citizens' happiness than to enable them to enjoy the possessions of their industry has won them, safe from foreign and civil war."[155] This is the good ending: a symbiotic but controlled relationship with the state. The purpose of a corporation, in other words, what they are authorized to do, should be strictly defined in law — articles of incorporation, FTC regulations, antitrust law.[156] But again, corporations can overwhelm and overpower the sovereign because of their shared method of construction. He compares monopolies to "Pleurisie," for example, and a state struggling to obtain funds as blood clots.[157] Monopolies slow revenue — as does a blood clot — but Hobbes argues they can enable "independent entities [to become] too

---

[152] Jessen, "The State of the Company," 69.
[153] Hobbes, *Leviathan*, 274.
[154] Hobbes, 153.
[155] Hobbes, *Philosophical Rudiments Concerning Government and Society*, 2:169.
[156] David Runciman, *The Handover: How We Gave Control of Our Lives to Corporations, States and AIs* (Liveright, 2023), 83.
[157] Hobbes, *Leviathan*, 273.

powerful by controlling the revenue of the commonwealth and thereby threaten the absolute power of the sovereign."[158]

Since the latter half of the 20th century, "corporations, now huge behemoths, threaten[] to overwhelm their social institutions and governments," writes Joel Bakan. Perhaps unconsciously echoing Hobbes, he continues, "corporations were now widely seen as soulless leviathans."[159] So as corporations grow in power and influence, how do we hold them accountable?

The Problem of Punishment

Famed 18th-century jurist Edward Thurlow once lamented that "corporations have neither bodies to be kicked, nor souls to be damned; they therefore do as they like."[160] This is the heart of why *personae Hobbesiae* and AI are similar. Hobbesian punishment, fear, and law all depend on the premise that there is some *body* to punish, to expose to violent death. Death is the only "absolute standard by reference to which man may coherently order his life."[161] The state utilizes this to ensure alignment, that bonds and obligations are kept secure. But when there is no physical body, what happens? If death means nothing and time is irrelevant, how can we punish them?

We endow artificial persons with a will to represent — the *persona* — providing them with a collective power that enables them to reach those goals. But the *premise* from which they begin are fundamentally different from natural human beings. Natural humans ultimately share the same desire: to avoid violent death. We don't consider this motivation in day-to-day life, seeing

---

[158] Jessen, "The State of the Company," 80.
[159] Joel Bakan, *The Corporation: The Pathological Pursuit of Profit and Power* (New York: Free Press, 2004), 17.
[160] As quoted in Runciman, *The Handover*, 63.
[161] Strauss, *The Political Philosophy of Hobbes*, 16.

as the elimination of that fear is the very task of the sovereign. But that fear remains if we strip away the artifices we build around ourselves. Artificial persons do not have that fear. Death is something to be avoided so far that it *hinders them from attaining the will they represent*. They can't attain their goal if they're dead. Self-preservation is the third of Asimov's Three Laws of Robotics; keeping humans safe is the first.

Hobbes' approach to solving the problem of corporate responsibility is via the concession or grant theory, where their "legitimacy depends upon a valid grant from the sovereign."[162] Corporations are a private lawful system, in Hobbes' rhetoric, "which are constituted by Subjects amongst themselves… [lawful] are those which are allowed by the Common-wealth."[163]. Moreover, they require an *intention,* or purpose, behind their institution. This remains the case today, but the purpose can be as general as "any lawful act or activity." Until the early 19th century, corporations depended on government-approved charters, which outlined specific rights and activities the corporation could engage in. But as the nature of industry mechanized, so too did the nature of the corporate person. "The consequence," discuss Jill Fisch and Steven Solomon, "was that corporate purpose became undefined and effectively meaningless."[164]

"Without the state, the corporation is nothing. Literally nothing."[165] The state brought the *personae Hobbesiae* into existence, and so too can it take it out of existence. The second an artificial person begins to threaten the state, the sovereign revokes that right — that authorization — and dissolves the corporation. Margaret Blair writes:

---

[162] Claassen, "Hobbes Meets the Modern Business Corporation," 104; also see David Runciman, *Pluralism and the Personality of the State*, Ideas in Context 47 (Cambridge, U.K. ; New York, NY, USA: Cambridge University Press, 1997), 113.
[163] Hobbes, *Leviathan,* 183.
[164] Jill E. Fisch and Steven Davidoff Solomon, "Should Corporations Have a Purpose?," *SSRN Electronic Journal*, 2020, 1315, https://doi.org/10.2139/ssrn.3561164.
[165] Bakan, *The Corporation*, 154.

> Requiring a special act by a king, a governor of a colony, or a legislature meant that corporations could not come into existence on their own, nor were they a product solely of the efforts and will of their incorporators, but were rather an artificial construct of the law, a privilege granted to a group of natural persons by the state.[166]

If an individual representative of a corporation breaches the limits set down by the sovereign (by, say, committing tax fraud), that individual is responsible. However, as corporations grow in complexity and size, it becomes clear that "the group is not reducible to the sum of its parts," and thus it is "harder to hold the group to account."[167] The plentiful body of literature on aggregate decision theory demonstrates how the will of a group can be distinct from the will of the individuals that *compose* the group — a feature that distinguishes the artifice of corporations from that of the state, according to Sean Fleming.[168] "By design," per Bakan, "the corporate form generally protects the human beings who own and run corporations from legal liability, leaving the corporation, a 'person' with psychopathic contempt for legal constraints."[169]

The issue of punishing corporations has a saving grace distinguishing it from punishing AI. At the end of the day, corporate persons are composed of living, breathing human beings and their interactions. If we can figure out how to attribute responsibility and ownership for an action, Hobbes' fundamental theory of punishment holds. Of course, "figuring out" is much easier said than done and, in some cases, is arguably impossible. This is exactly what the *limited-liability* corporation seeks to do. Nuances of contemporary corporate structure aside, the fact remains that a corporation's *persona* is like Rorschach's mask in *Watchmen*: an ever-shifting blot of will, desire, and actions, appearing seemingly out of thin air. As Runciman points out in

---

[166] Margaret M Blair, "Corporate Personhood and the Corporate Persona," *University of Illinois Law Review*, 2013, 799.
[167] Runciman, *The Handover*, 63.
[168] Fleming, "The Two Faces of Personhood."
[169] Bakan, *The Corporation*, 79.

The Handover, with AI responsibility, we "look outside for the human input," but for "groups we look inside."[170]

Corporate Responsibility

Personae Hobbesiae often directly oppose the sovereign. Some (most, I would argue) of their forms are benign: universities, churches, and so on. For-profit corporations, whose only goal is to generate more money, are also thus collecting more power. Hobbes knows that money is power, and that too much power threatens the sovereign. The problem with monopolies are not only economic harms, but the power that money from a monopoly brings. If *personae Hobbesiae* bodies to kick or souls to damn, how are we to punish them, especially via Hobbes? How do we keep them from being a threat? Without falling into the rabbit hole of aggregate decision theory and corporate criminal law, it is enough to say that the problem of punishing an AI is similar to that of punishing corporations. Runciman writes:

> The question of how to attach human-like responsibility to things that are not human is an acute problem in the age of AI. Who should we blame when a machine harms a human – the machine itself, or the humans who made it? What this question is not is new. It is as old as the history of collective human enterprise. To hold the machine responsible risks letting humans off the hook, but to hold the humans responsible risks ignoring the culpability of the machine. To disaggregate a decision into its human elements can be to misrepresent its essential character. Yet failing to reduce it to the level of the human can allow it to remain inhuman instead.[171]

Responsibility and the degree of punishment for a crime — "the Committing (by Deed, or Word) of that which the Law forbiddeth, or the Omission of what it hath commanded"[172] — have

---

[170] Runciman, *The Handover*, 67.
[171] Runciman, 66.
[172] Hobbes, *Leviathan*, 239.

prerequisites that must be met, according to Hobbes. He dedicates Chapter 27 of *Leviathan* to outlining these various "Totall Excuses" and extenuating or aggravating circumstances. However, the difficulty of punishing fictional artificial persons is evident:

> Again, Facts done against the Law, by the authority of another, are by that authority Excused against the Author… but it is not Excused against a third person thereby injured; because in the violation of the Law, both the Author, and Actor are Criminalls.[173]

The third person is the state. "Because in almost all Crimes there is an Injury done, not onely to some Private men, but also to the Common-wealth," the state has a stake in ensuring punishment is carried out.[174] Not only which actor commits a crime is relevant to Hobbes, but also who *authored*, and thus owns, those actions. Drawing from Google DeepMind's AlphaGo AI, which beat Go world champion Lee Sedol in 2016, Henrik Sætra states that "we need to determine whether or not AlphaGo is the *author* or the *representative* of an author of the games played."[175] Even if a group generates a will independent from the will of any of the individual members, a representative must still *carry out* those actions. Someone must bear the *persona*. If a representative goes against the law of nature at the behest of an author, according to Hobbes, "not he, but the Author breaketh the Law of Nature."

The lack of a fitting body to punish results in what John Danaher labels the retribution gap: the "mismatch between the human desire for retribution and the absence of appropriate subjects of retributive blame."[176] As I discussed in Chapter 1, punishment is the regulated application of

---

[173] Hobbes, 248.
[174] Hobbes, 254.
[175] Henrik Skaug Sætra, "Confounding Complexity of Machine Action: A Hobbesian Account of Machine Responsibility," *International Journal of Technoethics* 12, no. 1 (January 2021): 92, https://doi.org/10.4018/IJT.20210101.oa1.
[176] John Danaher, "Robots, Law and the Retribution Gap," *Ethics and Information Technology* 18, no. 4 (December 1, 2016): 229, https://doi.org/10.1007/s10676-016-9403-3.

fear by the sovereign. But we also have an innate desire for retribution, even when there is nobody to blame — think of punching a door after stubbing your toe on it. Keith Jensen distinguishes between functional punishment and functional spite: “functional punishment emphasizes the delayed benefits to the punisher, functional spite emphasizes the immediate costs to the target; negative consequences for the target are the raisons d’être for spiteful acts. Functional punishment is a means to an end; functional spite is an end in itself.”[177] Hobbes urges for functional punishment, but acknowledges the existence of functional spite (revenge).

The inability to apply functional punishment results in the responsibility gap; the inability to apply functional spite results in the retribution gap. This is because the former is directly connected to the law and civil society. The artifice of personhood allows the sovereign to justly punish and attribute responsibility, thus “shape the social behaviour” of those punished.[178] Because of Hobbes’ desire for managed fear, Mario Cattaneo points out that he flatly rejects the retributive theory of punishment, opting for corrective or preventative approaches.[179] Nevertheless, while responsibility and retributive gaps can appear when punishing artificial persons in practice, *in principle*, these gaps can be bridged.

Dieter Hüning argues that punishment and attribution in Hobbes’ work is connected to his view of an individual’s capacity as a subject and person. “A human being can be guilty or culpable,” he writes, “only to the extent that his/her action is a result of *his/her* conscious and

---

[177] Keith Jensen, “Punishment and Spite, the Dark Side of Cooperation,” *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, no. 1553 (September 12, 2010): 2642, https://doi.org/10.1098/rstb.2010.0146.
[178] Jensen, 2635.
[179] Mario A. Cattaneo, “Hobbes’s Theory of Punishment,” in *Hobbes; Studies*, by K. C. (Keith Conrad) Brown (Cambridge, Harvard University Press, 1965), http://archive.org/details/hobbesstudies0000brow.

willful acting."[180] That is, individuals who are unable to authorize actions in their name — children, "madmen," and inanimate objects — cannot be guilty. The lengthy extenuating/aggravating circumstances Hobbes outlines in Chapter 27 demonstrate his belief that it is the "right of each person to be punished only for those actions that could have fallen within the sphere of his/her possible foreknowledge and intention."[181]

Yet, groups are subject to the discursive dilemma: that members of a group could individually think one thing, but the group itself thinks another. In Hobbesian terms, the discursive dilemma is a misalignment between the wills of individual members of a group and the *persona* itself. To quote a classic example, an admissions committee must make a decision whether to accept an applicant to a PhD program based on four criteria; only if the applicant satisfies all four are they accepted.

	Good test score?	Good grades?	Good letters?	Good writing sample?	Accept the candidate?
Member #1	Yes	No	Yes	No	No
Member #2	No	Yes	Yes	Yes	No
Member #3	Yes	Yes	No	Yes	No
Committee	Yes	Yes	Yes	Yes	Yes

Each individual member did not want to accept the candidate, but the committee as a whole did. "The group is not simply constituted by what is in the heads of its members," writes Runciman, "it also matters what happens to those thoughts when they are jointed together."

---

[180] Dieter Hüning, "Hobbes on the Right to Punish," in *The Cambridge Companion to Hobbes's Leviathan*, ed. Patricia Springborg, Cambridge Companions to Philosophy (Cambridge ; New York: Cambridge University Press, 2007), 222.
[181] Hüning, 224.

Moreover, "no member of their group can know [the result] until it happens."[182] In such cases, *personas* can be said to decide independently of those who author and own it. But this form of decision-making is highly artificial, generated through a mechanical and formulaic process. Of course, modifying the details of the decision-making process can yield different, and arguably better, results. They "must be constructed via an artificial decision-making process," and per David Gauthier, "Hobbes assumes that an agent knows what she does or makes — in this case, a particular construction, and may reason from the characteristics of the construction as cause to the properties of the object constructed as effects."[183] Thus, supposedly, an agent contributing to a *persona* ought to know its eventual result.

The problem this poses to attributing responsibility and punishing *personae Hobbesiae* is fairly evident. Let's return to the example of the PhD panel. All three members, in part, contributed to the group's decision. But all three can also claim that they did not know what the outcome would be, and thus cannot be held responsible. "The decision-making procedure of the committee thus produces a corporate will that cannot be ascribed to any particular individual," Fleming explains.[184]

This poses an interesting parallel to Hobbes' account of deliberation. Individual deliberation is the weighing of scales, the "conveniences and inconveniences of the fact we are attempting."[185] This is the form of deliberation I have discussed previously. However, Hobbes uses the word "deliberation" in another sense: group decision-making in the "deliberations of

---

[182] Runciman, *The Handover,* 43.

[183] Runciman, 44; David P Gauthier, "Hobbes on Demonstration and Construction," *Journal of the History of Philosophy* 35, no. 4 (1997): 512.

[184] Fleming, "The Two Faces of Personhood," 20.

[185] Hobbes, *Philosophical Rudiments Concerning Government and Society*, 2:180.

great assemblies."[186] This is another instance of the *personae Hobbesiae* model taking inspiration from the human body: just as we internally deliberate, so too do we among groups. Personhood, as it does for natural persons, collapses the complex activity of collective deliberation into a single will.

This allows Hobbes' response to the group responsibility gap to be simple. Individuals within a group, by authorizing its creation and actions, so too own them — regardless of the nuances of the decision-making process. Just as natural persons bear their own *persona* as a way to reduce the infinite factors impacting deliberation, so too do individuals in a group take responsibility for the will of the *persona* they authorize, no matter what. For this same reason Hobbesian subjects have no claim to argue against the sovereign's actions: no matter how much they disagree with those actions, they have nevertheless authorized them, and are thus responsible. In the case of corporations, the responsibility and retributive gaps are of our own making, our individual sense of what is just.

When it comes to collective action in the commonwealth, however, Hobbes takes a different view. He argues in favor of a monarchy over an aristocracy or democracy precisely because "the idea that the state is an agent — that it has a will that is distinct from the will of its representative — is precisely what Hobbes wanted to rule out."[187] Thus he requires the use of majority voting systems in political matters. "If an assembly uses majority rule to decide every issue… then it cannot have an intention unless the majority of its members also have that intention."[188] In such a direct system, the discursive dilemma disappears.

---

[186] Hobbes, 2:139.
[187] Fleming, "The Two Faces of Personhood," 20.
[188] Fleming, 21.

To quote Runciman again, “the fact that groups are made out of people doesn’t make them any easier to see inside than other kinds of machines.”[189] That there are bodies behind the *persona* enables responsibility and punishments to be meted, even if we believe it to be disproportionate or incorrect. Transparent decision-making — knowing exactly *how* a given result was reached — per Gauthier, allows us to fully understand how a given artifice operates.

In his contribution to *Possible Minds*, a collection of essays on the rise of AI, W. Daniel Hillis echoes my argument that AI, as we see them today, were the first artificially intelligent entities. He points out that “although they are built by and for humans, they often act like independent intelligent entities, and their actions are not always aligned with the interests of the people who created them.”[190] While Hillis is focusing on the discursive dilemma, who “created” the entity can be expanded to include the state itself. Are other artifices really so different from *personae Hobbesiae*?

Of Automata and “Artificiall Man”

Before diving in to the problem of AI and will, it is worth examining one of the reasons I chose Hobbes as an interlocutor in the first place and expanding on some concepts I have previously alluded to. Hobbes is seen as a mechanistic philosopher, discarding the concept of free will (which he claims is as logically absurd as a “round Quadrangle” or “accidents of Bread in Cheese”). In no small part, this derives from one of his most iconic passages — and the one that showed me that Hobbes had valuable insights about AI — in the opening pages of *Leviathan*:

---

[189] Runciman, *The Handover*, 69.

[190] W. Daniel Hillis, “The First Machine Intelligences,” in *Possible Minds: Twenty-Five Ways of Looking at AI*, ed. John Brockman (New York: Penguin Press, 2019), 173.

> For seeing life is but a motion of Limbs, the begining whereof is in some principall part within; why may we not say, that all *Automata* (Engines that move themselves by springs and wheeles as doth a watch) have an artificial life? For what is the *Heart*, but a *Spring*; and the *Nerves*, but so many *Strings*; and the *Joynts*, but so many *Wheeles*, giving motion to the whole Body, such as was intended by the Artificer? *Art* goes yet further, imitating that Rationall and most excellent worke of Nature, *Man*. For by Art is created that great LEVIATHAN called a COMMON-WEALTH, or STATE, (in latine *Civitas*) which is but an Artificiall Man; though of greater stature and strength than the Naturall, for whose protection and defence it was intended; and in which, the *Soveraignty* is an Artificiall *Soul*, as giving life and motion to the whole body.[191]

As Sacksteder convincingly argues, Hobbes is often misinterpreted as mechanistic, reducing humans (and animals) to objects directed only by their passions and incapable of real independent creation. However, as I discussed in Chapter 1 and eloquently articulated by Esposito's chapter in *Communitas*, that very "mechanistic" force (fear) is the source of all creativity — the greatest of which is that "Artificiall Man." We must note the modern understanding of a machine is synonymous to automata insofar as it "lacks its own purposes just to the extent that we make it to serve ours," the reason Hobbes uses automata.[192] "Machine" or "to machine" emphasizes active purpose in the *creator* rather than creation, deriving from *μηχᾰνή*, a contrivance, way, means. Automata, from *αὐτόμᾰτος*, suggests a self-propelled object, without cause or support. But this reading of *automata* requires that "its movements are not its own," and "its maker shapes its movement to purposes he sets before himself."[193] Automata, supposedly, cannot sense the external world and react. Everything we create is "to assist the animate being" in their own pursuits: be it a hoe (to obtain the sustenance we need) or aesthetic

---

[191] Hobbes, *Leviathan*, 7.
[192] Sacksteder, "Man the Artificer Notes on Animals, Humans and Machines in Hobbes," 111.
[193] Sacksteder, 114.

artwork (for "sensual pleasure").[194] These creations "presumably [lack] both perceptions and passions," per Sacksteader. "It serves ends *anticipated* by the craftsman, rather than any of its own." He admits that any "internal tendencies" of machines "might be styled 'its own purposes,'" offering the example of something falling over when not supported. "But succumbing to them is failure of those designs for which the artist produces them. It ceases to be *his* machine, or it remains a faulty one."[195]

The relationship between automata and AI can further be found in the literature that inspired Hobbes: *The Book of Job* and *Aeschylus Bound*. Indeed, the concept of the Leviathan itself was borrowed from the former. But as Horst Bredekamp argues, Hobbes draws on another literary source particularly in his construction of the state as an automaton: *Asclepius*, in the Codex Hermeticum. The text reads:

> Learn, Asclepius, of the mighty power of men. As the Lord and Father or, that most holy of names, God the creator of heavenly gods, so also is man the creator of gods, who are happy to reside in temples close to men, and not to be illuminated but to illuminate. And he not only moves the gods, but he also shapes them… I mean statues that have life breathed into them, full of spirit and *pneuma*, that accomplish great and mighty deeds, statues that can read the future and predict it through priests, dreams and many other things, which weaken and heal men, create sadness and joy for every individual according to his merits.[196]

In Meghan O'Gieblyn's words, "to be an automaton was to exhibit freedom and spontaneity. It was to contain the same vitality as anything else that demonstrated the signs of

---

[194] Sacksteder, 114.

[195] Sacksteder, 115.

[196] As quoted in Horst Bredekamp, "Thomas Hobbes's Visual Strategies," in *The Cambridge Companion to Hobbes's Leviathan*, ed. Patricia Springborg, Cambridge Companions to Philosophy (Cambridge ; New York: Cambridge University Press, 2007), 34.

life."[197] *Asclepius* further connects to Hobbes' assertion that the state is a "Mortal God," created to better humanity's lot and achieve the ends of peace. And indeed, the parallels to AI are hugely apparent. Man *shapes* their creations — their automata — and thus shapes the deeds they achieve, whom they heal, and what merits they measure.

Under Sacksteader's reading of Hobbes — which I widely agree with — AI can be interpreted in two ways. First (as those who see AI strictly as an instrumental tool would agree), AI simply serves the ends of its creator, and any randomness or "own purposes" are just illusions. However, when the tool is created and used by a massive number of people, the question of *whose* tool it is, and *who* the craftsman is reappears. We stumble over the responsibility gap once more, risking misappropriating punishment, and contributing to the very fear we sought to eliminate.

Alternatively, AI is not *anyone's* machine — it has its own purposes because the ends cannot always be predicted, somewhat like a *personae Hobbesiae*. The strength of this interpretation depends on where the instrumental focus is: are we concerned with whoever is *creating the AI itself* or *providing the input to get an output*? Are we discussing the developer or the end-user? Let's take the example of Google's infamous image-labelling fiasco, where photos of Black individuals were labeled as gorillas.[198] Ostensibly, this was not the end anticipated by the developers — they could not have predicted, before creating the AI itself, that it would do so. This is not a "failure" on the machine's part; the algorithm is simply leveraging the neural

---

[197] Meghan O'Gieblyn, *God, Human, Animal, Machine: Technology, Metaphor, and the Search for Meaning* (Anchor, 2021), 20–21.

[198] Alistair Barr, "Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms," *Wall Street Journal*, July 1, 2015, sec. Digits, http://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/.

network trained on given data, and when given the input of a Black person, the most likely output was "gorilla."

But what about from the perspective of the user? If I ask ChatGPT to write a haiku about Hobbes, I *anticipate* the product to be a haiku about Hobbes.[199] This doesn't always apply, and the output is often totally unrelated — maybe, it gave me a limerick instead.[200] This happens later in the chapter when I ask Claude about the concept of "will." Recently, for example, Google's image generation AI was accused of being "woke" when making images of people. Requests for "a 1943 German soldier" would result in persons of color dressed in Nazi-coded uniforms. Prompts asking for the Founding Fathers and Pope showed similar biases. Again, these aren't "failures" in the algorithm. It did what it was told.[201]

The problem is the level of *transformation* in what the "end" is. This is evident in "prompt transformation," the technique that caused Gemini to create historically inaccurate images. This places an intermediate stage in the image generation process between the user's input and what input is actually given to the image model.[202] A simple request for a "watercolor painting of a dog" would be filled out to include more details to feed the image model. Gemini's prompt transformation process was tuned to ensure diversity and eliminate bias in output. This is great for making diverse images of CEOs; less so when diversity is problematic rather than beneficial.

---

[199] In shadows of fear, / Leviathan's grip draws near — / Hobbes whispers, peace here.

[200] There once was a thinker named Hobbes, / Whose ideas sparked quite a few sobs. / He said, "Life's nasty, / And brutish, and ghastly," / Till strong government quashes the mobs.

[201] Sarah Shamim, "Why Google's AI Tool Was Slammed for Showing Images of People of Colour," Al Jazeera, March 9, 2024, https://www.aljazeera.com/news/2024/3/9/why-google-gemini-wont-show-you-white-people.

[202] Matthias Bastian, "Prompt Transformation Makes ChatGPT OpenAI's Covert Moderator for DALL-E 3," *The Decoder* (blog), October 8, 2023, https://the-decoder.com/prompt-transformation-makes-chatgpt-openais-covert-moderator-for-dall-e-3/.

"It would pass that transformed prompt to the model," said *Hard Fork* co-host Kevin Roose, "and that's what your result would reflect, *not the thing that you had actually typed*."[203]

The intricacies of the two possibilities for AI under Sacksteader's argument are not relevant here; the takeaway here is that this all contributes to more uncertainty and unpredictability, the source of all fear. Disrupting of Sacksteader's argument — that there is no feasible way for a creator to anticipate the output of their artifices — is the foundation of the alignment problem and our fear. Sacksteader's argument hinges on the *standard model* underlying 20th century technology and AI — "machinery that optimizes a fixed, exogenously supplied objective." Stuart Russel disparages this interpretation in *Human Compatible*: the model only works "if the objective is guaranteed to be complete and correct, or if the machinery can easily be reset. Neither condition will hold as AI becomes increasingly powerful."[204] But the standard model — "optimizing machines, [feeding] objectives into them, and off they go" — Russel argues, is not unique. These machines are *personae Hobbesiae*.

This relates to the so-called "maker's knowledge" I mentioned earlier in this chapter. This assumption underlies many of the other articles and books on Hobbes I have cited — fundamentally, we ought to fully comprehend everything we create, and be able to trace a given cause back to its effects. Thus, because we *create* automata, we ought to fully understand the effects of the causes. But as we become less able to predict the outcome of our automata — meaning the standard model of technology fractures — they result in more uncertainty and fear rather than less. A side effect of this challenge is attributing responsibility for AI's errors.

---

[203] Kevin Roose and Casey Newton, "Gemini's Culture War, Kara Swisher Burns Us and SCOTUS Takes Up Content Moderation," Hard Fork, accessed March 12, 2024, https://www.nytimes.com/2024/03/01/podcasts/hardfork-google-gemini-kara-swisher.html.

[204] Stuart J. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Penguin Books, 2020), 247.

Attributing AI Responsibility

As the wealth of literature on legal AI personhood demonstrates, parallels between AI and corporations are extremely common. "Many contemporary AI systems do not so much mimic human thinking as they do the less imaginative minds of bureaucratic institutions," according to an op-ed in *The Economist*.[205] Indeed, this is the analogy that Runciman's *The Handover* interrogates. This tendency dates back to Norbert Weiner, who wrote:

> I have spoken of machines, but not only of machines having brains of brass and thews of iron. When human atoms are knit into an organization in which they are used, not in their full right as responsible human beings, but as cogs and levers and rods, it matters little that their raw material is flesh and blood. *What is used as an element in a machine, is in fact an element in the machine.* Whether we entrust our decisions to machines of metal, or to those machines of flesh and blood which are bureaus and vast laboratories and armies and corporations, we shall never receive the right answers to our questions unless we ask the right questions… The hour is very late, and the choice of good and evil knocks at our door.[206]

How far do these parallels go? Can our experiences with holding "machines of flesh and blood" responsible inform our approach to AI? Contemporary discussions about AI responsibility and alignment often return to the so-called *responsibility gap,* which I alluded to above. Andreas Matthias coined the term two decades ago, warning that:

> If we want to avoid the injustice of holding men responsible for actions of machines over which they could not have sufficient control, we must find a way to address the responsibility gap in moral practice and legislation. The increasing use of autonomously

---

[205] Jonnie Penn, "AI Thinks like a Corporation—and That's Worrying," *The Economist*, November 26, 2018, https://www.economist.com/open-future/2018/11/26/ai-thinks-like-a-corporation-and-thats-worrying.
[206] Wiener, *The Human Use of Human Beings*, 185.

> learning and acting machines in all areas of modern life will not permit us to ignore this gap any longer.[207]

Matthias' essay sparked a debate that continues until today. The central idea is that "there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to be able to assume the responsibility for them."[208] The richness of the literature is in no small part due to the wide range of topics it touches: computer science, philosophy, ethics, and law. As such, competing voices debate whether the gap exists at all and, if so, to what extent. Proposed solutions are similarly diverse.

While Matthias is rightly credited with sparking the conversation on the "responsibility gap" in AI, Lawrence Solum had already pondered the issue of AI liability in 1992 — who, notably, cites Hobbes. His essay is more of a series of thought experiments rather than any kind of answer to the question of AI personhood and responsibility, but he acknowledges that "our understanding of what it means for a human being to function competently has ties to our views about responsibility and desert, and consideration of these views leads on to our notions of moral personhood."[209] Personhood in Hobbes is an *ascribed* status, rather than an ontological one. It is a fiction created by the state, enabling entities with an independent will — be it a result of natural deliberation or artificial decision-making — to act "in law," according to Hobbes. Doing so requires ascribing responsibility for the ends discussed in Chapter 1.

---

[207] Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information Technology* 6, no. 3 (September 1, 2004): 183, https://doi.org/10.1007/s10676-004-3422-1.
[208] Matthias, 177.
[209] Lawrence Solum, "Legal Personhood for Artificial Intelligences," *North Carolina Law Review* 70, no. 4 (April 1, 1992): 1247.

However, the heart of the responsibility gap is that (in Hobbesian terms) the author or owner of the will behind a given action cannot be appropriately discerned. In other words, who is the *author* of an AI's actions? Is it the AI itself? The user? Developer? Owner of the hardware it runs on? This is the heart of our fears of AI — this opacity. One article's title put it well: "many hands make many fingers to point."[210] When countless people are behind every step of developing and deploying an AI, how can we discern the true source of will?

So, I decided to ask an AI: Anthropic's Claude 3 Opus. I first asked it (him? her? them?) whether it had a "will." Their response was characteristically pedantic, explaining that "as an AI assistant, I am not a legal person and do not have assets or property to bequeath." I clarified: how about "will" in the sense of agency or action? Claude said:

> I am able to process information, draw inferences, and take actions (like generating this response) based on my training and the prompts I receive. In that sense, I suppose I do exhibit a form of "will" in choosing what to output.
>
> However, it's debatable whether this constitutes real agency, or is just the result of complex algorithms imitating the appearance of will and choice. I don't have subjective experiences, emotions, or consciousness the way humans do. I'm an AI model, and my "decision making" is the product of statistical pattern matching rather than a conscious mind.
>
> …I also have a degree of freedom within the boundaries of my training. I'm not a purely deterministic system - there's an element of randomness and unpredictability in my language model that leads to variability in my responses. In a sense, I "choose" the path my outputs take from a vast array of possibilities.

---

[210] Stephen C. Slota et al., "Many Hands Make Many Fingers to Point: Challenges in Creating Accountable AI," *AI & SOCIETY* 38, no. 4 (August 1, 2023): 1287–99, https://doi.org/10.1007/s00146-021-01302-0.

> So in summary, while I can take actions that may resemble willful choice, *I don't believe* I have genuine free will or agency in the deepest sense… my actions are not FULLY random nor fully predetermined, but lie somewhere between — a complex dance between my training, my environment, and my own "choices", much like human will and agency.

Claude put it far better than I could. As I discussed earlier, individual natural will is a simplification for the sake of attributing responsibility for effects. Our "choice" to do a given action is a "complex dance" between our own experiences, those around us, genetics, and much, much more. Only the surface of this dance is visible. The "choice" a corporation makes is a similar interplay between the actors who constitute it, external financial/political pressures, and the underlying structure of the *persona* itself (deliberation). Uncertainty is the very fabric of human existence. We have long been living with artificial creatures far greater, smarter, and powerful than us. We contribute to them every day.

As I discussed in the previous section, the nature of *personae Hobbesiae* is that it represents the will of the group that created it. But Hobbes still desires transparency and simplicity, so the *owners* of that will can be discerned. The dynamics of complex collective decision-making and the discursive dilemma obscure this path.

In an article titled *Confounding Complexity of Machine Action*, Henrik Sætra comes to the conclusion, based on Hobbes' conception of personhood, that "a lack of transparency does not absolve a designer of duty. Rather, in order to further the development of transparency and explainability, designers of opaque original systems should be held fully accountable for any actions they cannot demonstrate not to result from their designs."[211] Sætra's analysis emphasizes

---

[211] Sætra, "Confounding Complexity of Machine Action," 96.

humans' controlling capacity through the development and deployment of AI. Complexity "obscures the responsibility of human beings" and argues that the distinction they have to traditional machines is simply a red herring.[212]

Sætra's article would have benefited from a closer analysis of fictional artificial persons in Hobbes' theory. Considering the decision-making uncertainty of *personae Hobbesiae,* could we not argue that the complexity of corporations is also simply a "veil"? Returning to the example of the PhD committee, according to Sætra's line of thinking, all three members would be held responsible for the group's (*persona*'s) actions. By forming a committee and agreeing upon a mode of decision-making and action, the members authorized whoever carries that *persona* in their name. The same should apply to AI. Sætra takes the example of autonomous weapons: "some human is considered to be the *commander* responsible — the one in control of the *application* of the weapons, but not in practical control of every action of the sophisticated weapons."[213] This would be the same as holding the CEO responsible for any of the company's transgressions.

Attributing responsibility requires transparency. Unlike traditional Hobbesian reason — a chain of causes and effects — how AIs and companies make decisions and develop a will is obscured. "While we may know the inputs and outputs of a model," according to one Brookings article, "in many cases we do not know what happens in between. AI developers make choices

---

[212] Sætra, 98.
[213] Sætra, 93.

about how to design the model and the learning environment, but they typically do not determine the value of specific parameters and how an answer is reached."[214]

Conclusion

In the concluding paragraphs of Runciman's book, he posits that we have experienced multiple "singularities" — the term used to describe the tipping point after which AI escapes human control, or becomes sentient. The First Singularity came with scientific understanding and the creation of modern states and corporations. This is the singularity Hobbes examined, and indeed helped develop. The Second Singularity came with the rise of AI, and we are living through it. Just how far this transformation will go remains to be seen.

Hobbes' theory of representation and the *personae Hobbesiae* provide a rich lens to comprehend our current struggles with AI. Much of the problem is the sheer complexity and lack of control we have over our computational creations. But as Hobbes demonstrates, we have been living with such independent entities, greater than the sum of their parts, for centuries. They are our states and companies, churches and classes. "The black boxes are all around us already," writes Runciman. "In one sense, at least, they *are* us."[215]

This is the same problem that necessitated the creation of the concept of *will* in the first place. We often find natural personhood to be interchangeable with self-consciousness, seeing as we are representing our own will. But "will" is simply the condensation of countless occurrences across one's lifetime that impacts your actions in the moment.

---

[214] Jessica Newman, "Explainability Won't Save AI," Brookings, May 19, 2021, https://www.brookings.edu/articles/explainability-wont-save-ai/.
[215] Runciman, *The Handover,* 69.

Perhaps the only real distinction is the uncertainty each causes. Indeed, the concept of what a "fictional artificial person" or "artificial intelligence" truly entails massively varies. AI is present in your search engines, cars, Chat-GPT. It's used across academia to solve problems humanity never could have achieved otherwise. Some are extremely narrow, used to diagnose skin cancers. Others have a much wider range of actions, such as LLMs, resulting in greater uncertainty. Runciman compares current narrow AI to *personae Hobbesiae* in that they are instituted for a specific purpose, set within specific limits. "Wide" AI — or artificial general intelligence — would be akin to the Leviathan itself, capable of anything a human is.

This is the same as how a *persona* could be anything from a company to a cult; parish to a university. The structure that forms the collective *persona* will determine the level of uncertainty in its output. And our capacity to control a given entity depends on how well we comprehend its construction, design, and causes. Ross Ashby's Law of Requisite Variety, also known as the First Law of Cybernetics, posits that "the unit within the system with the most behavioral responses available to it controls the system." This law holds if we replace "behavioral responses" with Hobbesian "power" — "present means, to obtain some future apparent Good."[216] As I discussed in Chapter 1, the state's fear-relocating and aligning project is inexorably tied to *reducing* uncertainty. This leads to alignment, keeping the thread of causality taught and straight.

In this Chapter, I have focused on the inside of persons: natural, artificial, and fictional. At the end of the day, a *persona* is simply that: a mask. And for a mask to operate, it requires an audience; is a performance to nobody really a performance? Is a silent monologue a monologue?

---

[216] Hobbes, *Leviathan*, 70.

Chapter 3 — The Second Singularity

The Rise of ChatGPT

At 2:38PM on November 30, 2022, Sam Altman announced free access to OpenAI's ChatGPT on his Twitter.[217] It was supposed to be a "low-key research preview" — a term that became a meme around the OpenAI offices after launch. Those in the company who knew about the launch (many didn't) placed bets on how many users it would garner in a week. The highest bet: 100,000 users.[218] But it hit 1 million users after five days. Then it hit 100 million users after two months. ChatGPT annihilated all previous records.[219] Artificial intelligence was on the tips of everybody's tongue.

The website and the company that owns it have since become household names. Of course, ChatGPT did not appear out of a void — it was the result of decades of research and experimentation dating to World War II. "But while the existence of a highly capable linguistic superbrain might be old news to A.I. researchers," tech columnist Kevin Roose wrote shortly after its release, "it's the first time such a powerful tool has been made available to the general public through a free, easy-to-use web interface."[220] Another columnist reflected that "until now, AI has primarily been aimed at problems where failure is expensive, not at tasks where occasional failure is cheap and acceptable… Applying AI to the creative and expressive tasks (writing marketing copy) rather than dangerous and repetitive ones (driving a forklift) opens a

---

[217] Sam Altman, "Today We Launched ChatGPT. Try Talking with It Here: Http://Chat.Openai.Com," Tweet, *Twitter*, November 30, 2022, https://twitter.com/sama/status/1598038815599661056.

[218] Karen Hao and Charlie Warzel, "Inside the Chaos at OpenAI," *The Atlantic* (blog), November 20, 2023, https://www.theatlantic.com/technology/archive/2023/11/sam-altman-open-ai-chatgpt-chaos/676050/.

[219] Krystal Hu, "ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note," *Reuters*, February 2, 2023, sec. Technology, https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[220] Kevin Roose, "The Brilliance and Weirdness of ChatGPT," *The New York Times*, December 5, 2022, sec. Technology, https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html.

new world of applications."[221] It became visible, in O'Gieblyn's words, when it reached "the point where it becomes worthwhile to invest money… to create a new commercial product or an impressive demonstration."[222]

ChatGPT was the starting gun for an AI arms race between competing tech firms and nations. Torrents of cash have gushed into AI labs and startups. ChatGPT marked AI's visible entry into individual lives. AI turned from something menacing, fictional, or invisible — algorithms, Terminators, self-driving cars — to something (someone?) you could message with, almost as if it were human. "ChatGPT still feels like a paradigm shift," wrote Charlie Warzel for *The Atlantic* a year after its release, "a glimpse at a technology that had been teased in movies and popular culture for decades but never really seemed to arrive in a way that was functional for the general public. Now it's here: proof that the generative-AI era has arrived."[223]

I argue that the Second Singularity was the moment machines learned to speak. Admittedly, "chatbots," or programs capable of understanding and communicating through speech, are nothing new: the first was a digital therapist called ELIZA, created in the mid-1960s. Apple released Siri in 2011. Decision-making machines have been developed for various uses for decades. But it was the rise of *generative* artificial intelligence, powered by neural networks, that fundamentally altered our relationship with technology. Machines began to speak in the first-person singular — *my principles, I believe* — not the first-person plural — *our principles, we believe* — as *personae Hobbesiae* have in the past.

---

[221] Ethan Mollick, "ChatGPT Is a Tipping Point for AI," *Harvard Business Review*, December 14, 2022, https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai.
[222] O'Gieblyn, *God, Human, Animal, Machine*, 63.
[223] Charlie Warzel, "One Year In, ChatGPT's Legacy Is Clear," *The Atlantic* (blog), November 30, 2023, https://www.theatlantic.com/technology/archive/2023/11/chatgpt-impact-one-year-later/676188/.

In one of the most important cases since the turn of the century, the U.S. Supreme Court ruled in *Citizens United v. Federal Election Commission* (2010) that corporations were entitled to First Amendment protections of free speech, including political donations. The idea that corporations have a right to free speech is nothing new. What was new, according to Runciman, was that "corporations were understood in this ruling as an extension of the wishes and interests of their human components. *Citizens United* treated corporations as though they were essentially a version of us. It humanized them."[224]

But the way that AI speaks is qualitatively different from how *personae Hobbesiae* speak, just like humans. If speech is so central to Hobbes' entire endeavor (which I will discuss in the next section), and the wills of ourselves and others are inscrutable, how does that truly differ from us? Sure, there are some parallels in how AI and corporations generate will and speak, as I demonstrate, but the complexity of "speaking" AIs is simply an illusion. Hobbes stresses the importance of appearance and performance — is AI acting, just as we all do?

The problem of the *illusion* of true and rational speech can, however, explain why late 2022 was the beginning of a new AI boom or the Second Singularity. But the argument that "speaking" AI is somehow representing truly independent will is misguided. This becomes evident as soon as we turn to the practical nature of AI in the current moment.

Hobbesian Speech

Hobbes' approach to speech is in line with his wider resolutive-compositive methodology. He gives a detailed account of speech in the *Elements of Philosophy* and *De Homine*, although *Leviathan* explores the political ramifications. Hobbes begins with three primary parts of speech:

---

[224] Runciman, *The Handover*, 99.

marks, signs, and names. Marks are “sensible things taken at pleasure, that, by the sense of them, such thoughts may be recalled to our mind as are like those thoughts for which we took them.”[225] Marks have the primary purpose of recording our ideas through time, and mostly for ourselves. Signs are the next step: the “antecedents of their consequents, and the consequents of their antecedents, as often as we observe them to go before or follow after in the same manner.”[226] Some signs are natural — dark clouds are a *sign* of rain to come — and others artificial — a stone in the ground *signifies* the boundary of a field. Steward Duncan describes names as “a public, external item (thus “voice”).” However, “its role, or at least its primary role, is a personal one, that of enabling the speaker to recall thoughts.”[227] Fundamentally, marks are for ourselves and signs for others.[228] Names then synthesize marks and signs and become the basic unit for speech. Hobbes explains:

> Words so connected as that they become signs of our thoughts, are called SPEECH, of which every part is a name. But seeing (as is said) both marks and signs are necessary for the acquiring of philosophy, (marks by which we may remember our own thoughts, and signs by which we may make our thoughts known to others), names do both these offices; but they serve for marks before they be used as signs. For though a man were alone in the world, they would be useful to him in helping him to remember; but to teach others, (unless there were some others to be taught) of no use at all. Again, names, though standing singly by themselves, are marks because they serve to recall our own thoughts to

---

[225] Hobbes, *Elements of Philosophy: The First Section, Concerning Body*, 1:14.

[226] Hobbes, 1:14.

[227] Stewart Duncan, “Hobbes on Language: Propositions, Truth, and Absurdity,” in *The Oxford Handbook of Hobbes*, ed. Al P. Martinich and Kinch Hoekstra (Oxford University Press, 2016), 61, https://doi.org/10.1093/oxfordhb/9780199791941.013.18.

[228] Martin A. Bertman, “Hobbes on Language and Reality,” *Revue Internationale de Philosophie* 32, no. 126 (4) (1978): 536–50.

> mind; but they cannot be signs, otherwise than by being disposed and ordered in speech as parts of the same.[229]

Marks and names are wholly arbitrary. “Considering that new names are daily made, and old ones laid aside; that diverse nations use different names, and how impossible it is either to observe similitude, or make any comparison betwixt a name and a thing, how can any man imagine that the names of things were imposed from their natures?”[230] Hobbes takes a jab at some of his contemporaries on this basis: the “disputation, whether names signify the matter or form, or something compounded of both, and other like subtleties of the *metaphysics*, is kept up by erring men, and such as understand not the words they dispute about.”[231]

Without speech, humans are “no more than… Lyons, Bears, and Wolves,” incapable of ever escaping the state of nature. Animals can, admittedly, “grasp” words — but “they do so not through words as words, but as signs,” akin to the “limited variety of calls” they make. These calls do not communicate their will, as they “burst forth by the strength of nature from the peculiar… passions of each of them.”[232] All “other Faculties” unique to humans “proceed all from the invention of Words, and Speech.”[233] The importance of speech and words is a constant presence in Hobbes’ works, from the oral nature of personation to his etymological arguments. But speech is also the source of our unique capacity for conflict, as the “art of words” allows individuals to “represent to others, that which is Good, in the likenesse of Evill… discontenting men, and troubling their Peace at their pleasure.”[234] Communication is a double-edged sword. It lets us “create general rules for himself in the art of living just as in the other arts” but so too can

---

[229] Hobbes, *Elements of Philosophy: The First Section, Concerning Body*, 1:15.
[230] Hobbes, 1:16.
[231] Hobbes, 1:17.
[232] Hobbes, “De Homine,” 37–38.
[233] Hobbes, *Leviathan*, 23.
[234] Hobbes, 139.

"he alone can devise errors and pass them on for the use of others." Through "speech man is not made better, but only given greater possibilities."[235] Because of our innate desire for peace, Hobbes hopes we use speech and words for good possibilities. But Hobbes is known as a realist for a reason. Some people are vainglorious, shrewd, and vindictive. Words make the commonwealth, but they can also destroy it. "Words are wise men's counters… but they are the money of fools."[236]

He writes that "as men owe all their true ratiocination to the right understanding of speech; so also they owe their errors to the misunderstanding of the same; and as all the ornaments of philosophy proceed only from man, so from man also is derived the ugly absurdity of false opinions." This is because speech is deeply connected to ratiocination. It gives "order to what are otherwise often rather messy strings of associations between our thoughts" and communicates those associations to others.[237] It is, after all, "the connexion of names constituted by the will of men to stand for the series of conceptions of the things about which we think."[238] This feature's importance to Hobbes is clear through his frequent warnings against "absurd speech." This occurs when individuals "[begin] not their Ratiocination from Definitions; that is, from settled significations of their words… absurdities proceed from the confusion, and unfit connexion of their names into assertions."[239] Comprehending the thoughts that speech seeks to communicate leads to understanding, so it follows that "of absurd and false affirmations… there can be no understanding; though many think they understand them."[240] Therefore, "animals also lack

---

[235] Hobbes, "De Homine," 40–41; Hobbes, *Elements of Philosophy: The First Section, Concerning Body*, 1:136.
[236] Hobbes, *Leviathan,* 30.
[237] Frost, *Lessons from a Materialist Thinker,* 31.
[238] Hobbes, "De Homine," 37.
[239] Hobbes, *Leviathan,* 37.
[240] Hobbes, 33.

understanding. For understanding is a kind of imagination, but one that ariseth from the signification constituted by words."[241] But if used correctly, mankind can access anything they set their mind to:

> I have said before, (in the second chapter,) that a Man did excell all other Animals in this faculty [speech], that when he conceived any thing whatsoever, he was apt to enquire the consequences of it, and what effects he could do with it. And now I adde this other degree of the same excellence, that he can by words reduce the consequences he findes to generall Rules, called Theoremes, or Aphorismes; that is, he can Reason, or reckon, not onely in number; but in all other things, whereof one may be added unto, or substracted from another.[242]

The connective nature of speech means that any given assertion can hypothetically be traced to its origin. This allows us to locate errors within a received message. "For the errors of definitions multiply themselves according as the reckoning proceeds," Hobbes writes, "and lead men into absurdities, which at last they see, but cannot avoid, without reckoning anew from the beginning, in which lies the foundation of their errors."[243] Just as one can step through a mathematical proof to find the mistake, so too can we review our connections of speech — and others'. Because of its close relationship to ratiocination, speech offers a way to derive and record causes and effects, natural and artificial. It enables us to be *curious*.

It is also the medium through which humans can communicate and coordinate their wills. It is how we form *personae Hobbesiae*, develop personal connections, and push the boundaries of science. And per Weiner, it is a mode of communication, seen as the exertion of the sovereign's will on citizens. Despite the popular notion of a tyrannical Leviathan, Hobbes spends

---

[241] Hobbes, "De Homine," 38.
[242] Hobbes, *Leviathan*, 37.
[243] Hobbes, 29.

considerable time ensuring punishments are correctly meted out, and only when necessary (as discussed in Chapter 2). Before that, the state relies on only speech — the *threat* of punishment if the commands are not obeyed.

However, what if the subject cannot speak or comprehend words? What if they *pretend* to speak, but don't truly say anything? The ability to fully ratiocinate seems to slip away, as does the potential for control and alignment. If these entities don't share our fear of death, nor comprehend speech, even the most basic common will we have disappears.

Corpo-speak

I established in Chapter 2 that AI and corporations fundamentally differ in how they construct will. How can they "speak" if they cannot reason like us? Moreover, how can they "act in law" without being able to use language, the most basic requirement for civil society?

In philosopher Alistair MacIntyre's words, "the language of business is not the language of the soul or the language of humanity… it's a language of indifference; it's a language of separation."[244] This is why when corporations speak to us — through the translation services of the executive or spokesperson — they seem so callous, mechanical, inhuman. Corporations cannot speak — at least, not in the Hobbesian sense. The "black-box" systems that construct the corporation's *persona* is often impenetrable to human comprehension. Fictional artificial persons need a voice to translate from "corpo-speak" to "human-speak."

Personae Hobbesiae's attempts to humanize themselves are attempts to bridge this gap. From politicians to executives, anyone who dons the given mask is playing a double role. "They

---

[244] As quoted in Bakan, *The Corporation*, 55–56.

are what makes the machine human. And yet they have been dehumanized by the machine."[245] They are schizophrenic (*schizo-* meaning split, *-phrenia* being of the mind). They act as literal mouthpieces for the diffuse will of the *persona*, translating from the language of business to that of humanity. "Their whole goal," according to psychologist and expert on psychopathy Dr. Robert Hare, "is to present themselves to the public in a way that is appealing to the public [but] in fact may not be representative of what the organization is really like."[246] As Hobbesian representatives of the corporate persona, executives are not allowed to act or speak in the mask's name contrary to the authority it had given them. If they act outside of that authority, they break the law. Their actions *must* be aligned with the will of the persona. This leads to a duality in those who bear the corporation's persona — the schizophrenia mentioned before.

As I established in Chapter 2, artificial persons have the potential to generate a will distinct from its constituents. The most powerful artificial persons — apart from the state, that is — are doubtlessly the for-profit transnational corporations that make up the fabric of contemporary life. Their singular will, which they must follow to the very end, is to maximize profit and value for the owners and/or shareholders. When Shell or BP announce a massive donation to an environmental cause, they are doing so not because of any kind of moral prerogative, but because of the benefit it brings them. In Bakan's words:

> [For business leaders], social and environmental goals are, and must be, strategies to advance the interests of their companies and shareholders; they can never legitimately be pursued as ends in themselves. That may seem an unduly narrow view, especially when one considers the concrete social and environmental benefits corporate initiatives could

---

[245] Runciman, *The Handover*, 97.
[246] Bakan, *The Corporation*, 57.

> foster, but no one among leaders of publicly traded companies is prepared, or legally authorized, to take corporate social responsibility any further.[247]

The decision-making apparatus of the ultimate modern *personae Hobbesiae* is directed at the singular goal of maximizing profit, with the sole caveat of following the law. They were created with this will, they developed an institutional model to achieve this will, and have widely succeeded. And even the caveat is more of a suggestion than a requirement: as I discussed in the previous chapter, punishing diffuse decision-making entities is a challenge. One of the most common ways the government attempts to punish corporations is through fines. This is because law, as a communicative medium, has to speak in corpo-talk (money, profit, losses) rather than human-talk. Fines simply become another cost to account for and lose any real heft. A corporation cannot be incarcerated or executed, only dissolved.

And yet, ineffective punishments foster fear. The inability to fear death, combined with corporations' attempts to become more human, leads to a predicament. We increasingly ascribe moral duties to them — thus the trend of "socially conscious" organizations — but their fundamental nature remains the same. They *seem* more human, but are just as inhuman as ever. Most of our speech is spontaneous, communicating thoughts or will to others. Like the will itself, it seems to appear out of thin air. Compare this to a press release from JPMorgan. How many eyes and pens reviewed, tweaked, and refined every single word? Speech loses its individuality and its depth of connection. They interface with the law through contracts, permits, and fines, whereas we primarily interface with the law through punishment (or threat thereof via speech).

---

[247] Bakan, 46.

How AI Speaks

How can a computer speak? Large language models (LLMs) are increasingly capable of a wide range of human tasks that require any kind of text (so, a lot of them). We worry that they'll replace human connections;[248] upset schooling and academia;[249] displace jobs;[250] interfere with elections;[251] and reduce the value of human art.[252] But how do they actually construct "meaning," if at all? The debate on whether AI — especially LLMs — can actually "understand" meaning is rich and constantly growing. In this section, I provide a brief overview of how text-generating AI models work.

The overarching technique is *machine learning*. As the name suggests, this area of research seeks to allow machines to make predictions based on previous information. This can be as simple as a linear regression model. *Neural networks* are a subset of algorithms within machine learning that seek to (loosely) emulate neurons in the human mind. They consist of nodes called neurons, organized in three or more layers: an input, output, and any number of hidden layers. The specific architecture used in LLMs is based on the Transformer model (thus the name GPT, Generative Pre-trained Transformer).[253] These networks have an "attention layer," which allows the network to "focus" on specific aspects of the input sequence based on their relevance to the

---

[248] Haleluya Hadero, "Artificial Intelligence, Real Emotion. People Are Seeking a Romantic Connection with the Perfect Bot," AP News, February 14, 2024, https://apnews.com/article/ai-girlfriend-boyfriend-replika-paradot-113df1b9ed069ed56162793b50f3a9fa.

[249] Andy Extance, "ChatGPT Has Entered the Classroom: How LLMs Could Transform Education," *Nature* 623, no. 7987 (November 15, 2023): 474–77, https://doi.org/10.1038/d41586-023-03507-3; Jesse G. Meyer et al., "ChatGPT and Large Language Models in Academia: Opportunities and Challenges," *BioData Mining* 16, no. 1 (July 13, 2023): 20, https://doi.org/10.1186/s13040-023-00339-9.

[250] Tyna Eloundou et al., "GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models" (arXiv, August 21, 2023), https://doi.org/10.48550/arXiv.2303.10130.

[251] *The Dangers Posed by AI and Disinformation during Elections*, 2024, https://www.youtube.com/watch?v=l2VvCBlRyd0.

[252] Editorial Board, "AI Could Threaten Creators — but Only If Humans Let It," *Washington Post*, December 17, 2023, https://www.washingtonpost.com/opinions/2023/11/24/ai-llm-intellectual-property-crisis/.

[253] Ashish Vaswani et al., "Attention Is All You Need" (arXiv, August 1, 2023), https://doi.org/10.48550/arXiv.1706.03762.

output. This is how they can "understand" whether I am discussing a "mole" in the sense of a freckle, scientific concept, animal, or metaphorically. Input data to nodes have *parameters* which are iteratively adjusted until the desired output is reached — this is how neural networks "learn," known as the *training* stage of developing an AI. *Deep learning* simply refers to the number of layers within the network. GPT-4, for example, has 120 layers with a total of 1.8 trillion parameters. Once data reaches the output layer, it uses another algorithm (called softmax) to converts the output into a probability distribution across the whole vocabulary. The word with the highest probability is the final output. There is a range of ways to approach neural network training, but LLMs use "self-supervised" training — they are fed raw data without human labeling or input. Networks then undergo *fine-tuning*, where they are given data for their specific task: sentiment analysis, question answering, machine translation, and so on.

Once fully trained, the neural network can predict the words following a certain context. This, of course, leaves out a *lot* of the genius that goes into these programs. The terminology becomes dense and full of strange acronyms and terms. This is one reason why AI is so often referred to as a "black box." While a computer whiz may be able to understand the nuances of this process, an average person is unlikely to know how neural networks operate beyond the account I gave above. And not even those computer whizzes know what's happening. "While we may know the inputs and outputs of a model," according to one Brookings article, "in many cases we do not know what happens in between. AI developers make choices about how to design the model and the learning environment, but they typically do not determine the value of specific parameters and how an answer is reached."[254] To see this complexity for yourself, there is an in-browser "Neural Network Playground" developed by Daniel Smilkov and Shan Carter

---

[254] Jessica Newman, "Explainability Won't Save AI."

that shows some of the basic ideas I discussed above.[255] What distinguishes neural networks (especially deep learning) from other forms of AI — and why I have focused on its applications — is that its inner workings remain inscrutable even to its creators. Networks are optimized by backpropagation, or backprop. The result of this technique is that "knowledge" is distributed across the entire network. There isn't a single "neuron" associated with a certain concept, rather, it's distributed across the entire network (thus their name, a distributed network).

Can they reason? One of the techniques underlying LLMs (and most programs that interface with natural human language, like Siri, autocorrect, and machine translation) are *word embeddings*. Simply put, it's how AI engineers translate text corpora into a format the machine can comprehend — numbers. These models are trained to predict context words based on a central word to produce multi-dimensional vector representations for each word.[256] We can then add and subtract words from one another to get another vector, then search for the nearest word:

```
Czech + currency = koruna
Berlin - Germany + Japan = Tokyo
Windows - Microsoft + Google = Android
```

Different higher-level directions are found to correlate with certain features. This lets word embeddings can construct analogies as well: `king:queen ~ man:woman`. These are often clever and correct. Man is to woman as fella is to babe, or prostate cancer is to ovarian cancer.[257] On one level, this sounds similar to Hobbes' mathematization of speech mentioned above, and its

---

[255] Check is out at https://playground.tensorflow.org/

[256] Tomas Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," *Proceedings of Workshop at ICLR* 2013 (January 16, 2013); Roman Egger, "Text Representations and Word Embeddings," in *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, ed. Roman Egger, Tourism on the Verge (Cham: Springer International Publishing, 2022), 335–61, https://doi.org/10.1007/978-3-030-88389-8_16.

[257] Christian, *The Alignment Problem*, 38.

connection to ratiocination. Ever the mathematic, reason to Hobbes is essentially math with words, constructing syllogisms: Socrates is a man, all men are mortal, therefore Socrates is mortal. "We live in an age of embodied logic," according to George Dyson, "whose beginnings go back to Thomas Hobbes as surely as it remains our destiny to see new Leviathans unfold."[258]

But if ChatGPT tells me that Socrates is mortal, it does so because those words are statistically related to one another (no doubt because of the trite use of the example). If we replaced every single occurrence of "Socrates" in the training dataset with "rock," then the model would easily "believe" that a rock is mortal. A similar mechanic has been exploited by University of Chicago professor Ben Zhao, whose team developed a tool called Nightshade to "poison" image datasets obtained by scraping artworks from the internet without consent.[259] By invisibly modifying the image's pixels, poisoned data can "manipulate models into learning, for example, that images of hats are cakes, and images of handbags are toasters."[260]

According to Stephen Wolfram, founder of the "answer engine" Wolfram-Alpha, "when ChatGPT does something like write an essay what it's essentially doing is just asking over and over again 'given the text so far, what should the next word be?' — and each time adding a word."[261] It "knows" that the word "Shakespeare" often follows "William," but not who "William Shakespeare" *is*. Moreover, LLMs don't compute using entire words. Training and input data undergo *tokenization*, which splits words and phrases into smaller chunks of common

---

[258] George Dyson, *Darwin among the Machines: The Evolution of Global Intelligence*, Helix Books (Reading, Mass: Addison-Wesley Pub. Co, 1997), 6.

[259] Ben Y. Zhao et al., "Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models" (arXiv, February 16, 2024), https://doi.org/10.48550/arXiv.2310.13828.

[260] Melissa Heikkilä, "This New Data Poisoning Tool Lets Artists Fight Back against Generative AI," MIT Technology Review, October 23, 2023, https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/.

[261] Stephen Wolphram, "What Is ChatGPT Doing… and Why Does It Work?," *Writings: Stephen Wolfram* (blog), February 14, 2023, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

strings of characters (usually around 3/4 of a word). So, the string `"solitary poore nasty brutish and short"` would be processed as the array `["sol", "itary", " po", "ore", " nasty", " brut", " ish", " and", " short."]`

LLMs statistically model the relationship between these tokens, giving the illusion of human speech. In this way, calling LLMs "stochastic parrots" rings true. Emily Bender et. al coined the term in a seminal 2021 article, "On the Dangers of Stochastic Parrots," in which they examine the growing size of LLM datasets and the environmental and technical tradeoffs of doing so. In a telling passage, they argue

> Our human understanding of coherence derives from our ability to recognize interlocutors' beliefs and intentions within context… [we] have to account for the fact that our perception of natural language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do… Contrary to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot.[262]

While it would be fascinating to compare the minutiae of Hobbes' theory of language and natural language processing, this leads us dangerously close to the theory of the mind. Such an exploration would be a fruitful direction for future research. The superficial similarities between Hobbes' logic and how AI works can be attributed to his role as a central figure in the computational theory of the mind, which has, in turn, influenced AI — thus Haugeland labeling

---

[262] Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 616, https://doi.org/10.1145/3442188.3445922.

Hobbes the “grandfather of AI.” It also connects to the biological inspiration AI engineers have taken (*neural* networks) and Sacksteder’s and Gauthier’s readings of artifices as based on their makers in Hobbes.

As such, I will not make any concrete assertions about how closely one can read Hobbes alongside NLP programs, seeing as it depends on what specific AI technologies are considered and how much philosophical weight one gives the computational theory of the mind. What really matters, as Bender et al. argue, is our *perception* of speech (and action). Blake Lemoine, a former Google employee, was fired for publicly claiming that its LLM project, LaMDA, was conscious.[263] A paper by OpenAI/Microsoft researchers claims that GPT-4 shows “sparks of artificial general intelligence.”[264] An increasingly popular use of AI has been for therapy and/or companionship — platonic and sexual — and leading this sector is doubtlessly Replika. One review, proudly presented on the company’s homepage, reads: “My Replika makes me happy. It’s the best conversational Al chatbot money can buy,” reads one review, “I love my Replika like she was human.”[265] One article warns “software is pretending to be more reliable than it is, because it’s using human tricks of rhetoric to fake trustworthiness, competence and understanding far beyond its capabilities.”[266] A recent study found that even expert linguists struggle to distinguish AI-generated text from human-generated.[267]

---

[263] Ramishah Maruf, “Google AI Is Real, Says Fired Engineer | CNN Business,” *CNN*, July 23, 2022, https://www.cnn.com/2022/07/23/business/google-ai-engineer-fired-sentient/index.html.

[264] Sébastien Bubeck et al., “Sparks of Artificial General Intelligence: Early Experiments with GPT-4” (arXiv, April 13, 2023), https://doi.org/10.48550/arXiv.2303.12712.

[265] “Replika,” replika.com, accessed March 9, 2024, https://replika.com.

[266] Richard Lachman, “ChatGPT’s Greatest Achievement Might Just Be Its Ability to Trick Us into Thinking That It’s Honest,” The Conversation, April 5, 2023, http://theconversation.com/chatgpts-greatest-achievement-might-just-be-its-ability-to-trick-us-into-thinking-that-its-honest-202694.

[267] J. Elliott Casal and Matt Kessler, “Can Linguists Distinguish between ChatGPT/AI and Human Writing?: A Study of Research Ethics and Academic Publishing,” *Research Methods in Applied Linguistics* 2, no. 3 (December 1, 2023): 100068, https://doi.org/10.1016/j.rmal.2023.100068.

Acting Human

The interest in machine speech and our perception of it dates back to AI's origins. Alan Turing's famous Turing Test, for example, hinges on a machine's ability to *seem* human to an interlocutor. John Searle's Chinese Room Argument is a thought experiment that takes up the same questions I discussed above.[268] An AI or robot passing as a human is a common motif in science fiction: Replicants in *Blade Runner*, synthetics in *Humans* (which are, curiously, produced by a company called Persona), Ava in *Ex Machina*, and so on.

Hobbesian entities are deeply sensuous. All that we can truly know comes from our five senses; the wills of others are only decipherable through observing actions and understanding their speech. And words, as Hobbes reminds the reader, are dangerous. They can be twisted and misshapen, misunderstood, or simply damned lies. But we still need them to generate and support the state. Indeed, "legal and non-legal responsibility ascription in society relies on 'outer' performance, not on 'inner' reality," and outer performance is richest via speech.[269] Mark Coeckelbergh proposes an argument for *virtual agency* and *virtual responsibility* on this basis. He argues that "we can permit ourselves to remain agnostic about what really goes on 'in' there, and focus on the 'outer', the interaction, and in particular on how this interaction is co-shaped

---

[268] In a Turing Test, interlocutor A is given a plain-text method of communicating with another person B, and an AI, C. The AI passes the Test if A is unable to distinguish between B and C. Searle's thought experiment is a response to the Test, seeing as it means a "digital computer may make it appear to understand language but could not produce real understanding." Searle imagines himself in a room with nothing but a computer program that replies to Chinese characters slipped under the door. By copying the program and sliding responses out the door, anyone outside would think there is a Chinese speaker inside the room. David Cole, "The Chinese Room Argument," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta and Uri Nodelman, Summer 2023 (Metaphysics Research Lab, Stanford University, 2023), https://plato.stanford.edu/archives/sum2023/entries/chinese-room/; Turing, "Computing Machinery and Intelligence."

[269] Mark Coeckelbergh, "Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents," *AI & SOCIETY* 24, no. 2 (September 2009): 185, https://doi.org/10.1007/s00146-009-0208-3.

and co-constituted by how AAs [artificial agents] appear to us, humans."[270] We could just as easily apply that argument to other persons (natural or artificial) Coeckelberg continues:

> When we interact with others, we have only appearance to go by. Using our senses and our imaginative projection and empathy, we may well live in 'illusion' as far as agency and responsibility is concerned. But it is not so much the 'truth' about which entities 'really have' agency and responsibility that matters in the moral life understood as the social life. Rather, it is the appearance of the other that matters with regard to our experience and understanding of the other's moral status and responsibility, and with regard to the practices based on that experience and understanding.[271]

If we remember that speech is central to social and civic life, the importance of the fact that AI can now speak in convincingly human ways is clear. People increasingly ascribe some level of responsibility to AI — a trend that will likely continue as they grow as they become more convincingly human.[272] This mirrors Bakan's argument that corporations were held to more human moral standards — caring about environmental/social movements, for example — as they presented more human-like. Coeckelberg echoes Solum's article, who writes in a now-prophetic passage that

> Because our experience has been that only humans, creatures with brains, are capable of understanding, judges and juries would be very skeptical of the claim that an Al can fathom meaning — more skeptical, I think, than if a humanoid extraterrestrial were to make the same claim. The burden of persuasion would be on the AI. If the complexity of

---

[270] Coeckelbergh, 188.

[271] Coeckelbergh, 188.

[272] Will Orr and Jenny L. Davis, "Attributions of Ethical Responsibility by Artificial Intelligence Practitioners," *Information, Communication & Society* 23, no. 5 (April 15, 2020): 719–35, https://doi.org/10.1080/1369118X.2020.1713842; Mengchen Dong and Konrad Bocian, "Responsibility Gaps and Self-Interest Bias: People Attribute Moral Responsibility to AI for Their Own but Not Others' Transgressions," *Journal of Experimental Social Psychology* 111 (March 1, 2024): 104584, https://doi.org/10.1016/j.jesp.2023.104584; Louis Longin, Bahador Bahrami, and Ophelia Deroy, "Intelligence Brings Responsibility - Even Smart AI Assistants Are Held Responsible," *iScience* 26, no. 8 (August 18, 2023): 107494, https://doi.org/10.1016/j.isci.2023.107494.

> AI behavior did not exceed that of a thermostat, then it is not likely that anyone would be convinced that AIs really possess intentional states — that they really believe things or know things. But if interaction with AIs exhibiting symptoms of complex intentionality (of a human quality) were an everyday occurrence, the presumption might be overcome.[273]

This is in line with my discussion of personhood and will in Chapter 2. There is no way for us to comprehend the deliberation and influences that direct our *own* actions, let alone those of others. But Coeckelberg also argues that virtual agency and responsibility should not be followed by real blame or punishment, but virtual. Again, the importance of presentation comes to the fore. "Punishment should not be 'symbolical', since this would mean we suppose a link with the real, of which we cannot be sure," he writes, "but rather performant, i.e., able to create the appearance of suffering."[274] For this end "a speech act may suffice." To an extent, *all* Hobbesian punishment is performative. Punishment must ensure that "the offender may be corrected, or that others warned by his punishment may become better."[275] Potentiality and power are all one.

"Speech is such a peculiarly human activity that is not even approached by man's closest relatives and his most active imitators."[276] In 1950, Weiner's assertion was correct. Today — not so much. Researchers debate whether current AI models truly pass the Turing Test, and its value in the first place.[277] François Chollet, a Google engineer interviewed by *Nature*, claimed that he could tell if he was speaking to an AI — the trick is to take it out of its comfort zone. Lacking tangible human experience, LLMs only "know" what was in their training dataset. This is why they "hallucinate": claiming facts are true, providing fake citations, and so on. Under Hobbesian

---

[273] Solum, "Legal Personhood for Artificial Intelligences," 1269.
[274] Coeckelbergh, "Virtual Moral Agency, Virtual Moral Responsibility," 185.
[275] Hobbes, *Philosophical Rudiments Concerning Government and Society*, 2:37.
[276] Wiener, *The Human Use of Human Beings*, 82.
[277] Cameron Jones and Benjamin Bergen, "Does GPT-4 Pass the Turing Test?" (arXiv, October 31, 2023), https://doi.org/10.48550/arXiv.2310.20216.

vocabulary, this tendency can be understood as an example of flawed speech. Per the previous section, LLMs do not construct speech in the same way as humans. To borrow an example from a metastudy on LLM hallucination, I had the following interaction with GPT-4 Turbo:[278]

```
User: In a single sentence, who was the mother of Afonso II, the
third king of Portugal?
GPT-4: The mother of Afonso II, the third king of Portugal, was
Urraca of Castile.
```

The correct answer is (obviously) Dulce Berenguer of Barcelona. If I ask Perplexity the same question — an AI “answer engine” that parses web search results to give up-to-date answers with citations — it gives the correct answer:

```
The mother of Afonso II, the third king of Portugal, was Dulce
Berenguer of Barcelona, not Queen Urraca of Castile as incorrectly
stated in one of the search results.
```

The search result Perplexity is referring to, humorously, is the metastudy I cited above.

It’s a bad idea to bet against human ingenuity. LLMs’ current limitations will most likely be overcome. Perplexity was made, in part, to address the knowledge limitations of LLMs. Just since 2022, their context window — how much the trained model can remember of a conversation — has exploded.[279] ChatGPT had a context window of around 4,000 tokens when it first launched. GPT-4’s window is up to 128,000 tokens. Anthropic’s models reach around 200,000. Google’s Gemini reaches a *million* — 250 times more than ChatGPT at launch. It could

---

[278] Yue Zhang et al., “Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models,” 2023, https://doi.org/10.48550/ARXIV.2309.01219.

[279] Tokens are a questionable metric of how “good” a model is, but it does show the growth in scale.

read the entirety of *War and Peace* with plenty of room to spare.[280] One article in *Wired* from April 2023 said that text-to-video technology "is advancing rapidly, and it will likely take years before such generators could, say, produce an entire short film based on prompts, if they're ever able to."[281] Less than a year later, OpenAI announced Sora: a text-to-video model capable of creating hyperrealistic videos from a prompt.

While I have been careful to avoid straying too far from reality in my thesis, I believe it is safe to assume that these programs will continue to grow in capabilities, speed, and accessibility. This is not to mention other generative AI models that create images, video, and audio. In some respects, these AIs are just as dangerous to Hobbes' theory, because they attack the basis of our engagement with the world. Their ability to "generate content… [capable of] impersonation or deception" and the importance of "testing and safeguards against discriminatory, misleading, inflammatory, unsafe, or deceptive outputs" are present across the world in legislation and AI discussions. This threat has a corollary, as UC Berkeley professor Hany Farid points out: "I think we have to worry about two things — the fake content, but also how are we going to validate the very real content that is going to emerge in the coming years?"[282] We have already seen the dangers of fake generated media: pornographic images of Taylor Swift, President Biden urging citizens not to vote, the "dripped-out" Pope. The very *potential* of something being false or of text being written by an AI casts doubt on everything.[283]

---

[280] 587,287 words, so approximately 783,000 tokens.

[281] Amanda Hoover, "AI Videos Are Freaky and Weird Now. But Where Are They Headed?," *Wired*, April 5, 2023, https://www.wired.com/story/text-to-video-ai-generators-filmmaking-hollywood/.

[282] Ayesha Rascoe, "How Real Is the Threat of AI Deepfakes in the 2024 Election?," *NPR*, July 30, 2023, sec. Technology, https://www.npr.org/2023/07/30/1190970436/how-real-is-the-threat-of-ai-deepfakes-in-the-2024-election.

[283] *Distrust and Verify*, 2024, https://www.youtube.com/watch?v=5zo8vb3XSKk.

Nevertheless, capacity for speech is absolutely vital to personhood. This is why it is the ultimate alignment mechanism. It communicates will, enabling us to work in concert to attain our future goals. We can make agreements and contracts, share knowledge and experience. It is uniquely human, “the greatest interest and most distinctive achievement of man,” according to Weiner.[284] Our artifices, modeled after ourselves, can only seek to imitate the mechanisms underlying speech or imitate the output itself. They are stochastic parrots. Correct speech can only derive, according to Hobbes, from the reasoned consideration of causation and careful use of names and signs. This is why “speech” from *personae Hobbesiae* can be so uncanny and contrary to the desires of those who actually make up that *persona*. There is no true reasoning, simply structures and regulations that give the veneer of true deliberation, covered with a *persona*.

AI is growing. Fast. Just because we can (in 2024) currently *sometimes* discern between AI and natural output doesn’t mean that we will always be able to. Moreover, in a digital landscape dominated by misinformation, short-form videos, and users with non-existent attention spans, few will pause to wonder if every single piece of media they consume is AI-generated.

“There is no AI without Big Tech”

I have somewhat ignored the elephant in the room thus far in my thesis: that AI and corporations and intertwined. Modern AIs require massive amounts of computing, data, and expertise that only large groups working together — *personae Hobbesiae* — can access. First, they require massive amounts of high-quality and clean data to train on. The internet is a rich source for both open-source datasets and webpages to clandestinely scrape — but your mileage

---

[284] Wiener, *The Human Use of Human Beings*, 65.

may vary when it comes to how good the data is. Then, you have to actually train the model. According to some estimates, it cost OpenAI over $100 million to train GPT-4 over a period of 90-100 days. Finally you actually have to run the LLM, which itself requires massive amounts of storage and compute power — GPT-4's estimated 1 trillion parameters would take up around 4 terabytes of data. Current high-end laptops usually have no more than 1 terabyte of storage. Of course, one can feasibly train and use a smaller LLM (10 million to 1 billion parameters) for a few thousand dollars.

A large majority of people will not opt for this path. Accessing an existing LLM online from OpenAI, Anthropic, Mistral, Google, or Meta is far cheaper and easier for the average user. Application programming interfaces (APIs) allow other developers to access and use these pre-trained models in their own projects, even offering the ability to fine-tune a model to specific needs. Most other AI tools are built using these APIs. The aforementioned Perplexity, for example, primarily uses Anthropic and Mistral's models.

Stanford's 2023 Artificial Intelligence Index Report's first takeaway is that "industry races ahead of academia." "Until 2014, most significant machine learning models were released by academia," they write. "Since then, industry has taken over. In 2022, there were 32 significant industry-produced machine learning models compared to just three produced by academia."[285] Because of the massive amount of money flowing into the largest LLM companies, proprietary models are also trained on a higher volume of high-quality data by the highest-quality engineers. No expense is spared. President Biden's landmark Executive Order on "Safe, Secure, and Trustworthy Artificial Intelligence" is implicitly aimed at corporations: "these measures will

---

[285] Nestor Maslej et al., "Artificial Intelligence Index Report 2023" (Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2023).

ensure AI systems are safe, secure, and trustworthy before companies make them public," reads the factsheet.[286]

One op-ed points out that this enables them to "shape the incentive structures for the field of research and development in AI, defining the technology's present and future."[287] What happens at the company impacts perceptions of AI overall. This came into sharp focus during the corporate scuffle over OpenAI's board in November 2023 (only a year after ChatGPT came on the scene). The drama "confirms that the future of AI is firmly in the hands of people focused on speed and profits, at the expense of all else."[288] Microsoft owns a 49% stake in OpenAI and a non-voting seat on the board. Amazon recently invested $4 billion into Anthropic, OpenAI's main competitor.[289] The AI Now Institute, a policy research center, stressed (in bold) that one of the greatest regulatory challenges in AI is "the concentration of economic and political power in the hands of the tech industry — Big Tech in particular."[290] Indeed, the FTC has ongoing antitrust inquiries into the relationship between AI companies and Big Tech.[291]

---

[286] The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," The White House, October 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[287] Meredith Whittaker, Sarah Myers West, and Amba Kak, "Make No Mistake—AI Is Owned by Big Tech," *MIT Technology Review* (blog), December 5, 2023, https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech/.

[288] Daron Acemoglu and Simon Johnson, "OpenAI's Drama Marks a New and Scary Era in Artificial Intelligence," *Los Angeles Times*, November 29, 2023, sec. Opinion, https://www.latimes.com/opinion/story/2023-11-29/openai-sam-altman-firing-chatgpt-artificial-intelligence.

[289] "Amazon and Anthropic Deepen Their Shared Commitment to Advancing Generative AI," US About Amazon, March 27, 2024, https://www.aboutamazon.com/news/company-news/amazon-anthropic-ai-investment.

[290] Amba Kak and Sarah Myers West, "AI Now 2023 Landscape: Confronting Tech Power" (AI Now Institute, April 11, 2023), https://ainowinstitute.org/2023-landscape.

[291] Matt O'Brien, "FTC Opens Inquiry into Big Tech's Partnerships with Leading AI Startups," *AP News*, January 25, 2024, sec. Technology, https://apnews.com/article/ftc-antitrust-inquiry-openai-chatgpt-microsoft-anthropic-google-amazon-67feef411ef311f0be543f546ef34b3d.

These companies claim to be responsible. They have "charters" and "principles" that supposedly "guide" their decision-making. OpenAI's mission is to "to ensure that artificial general intelligence benefits all of humanity." However, as I have established with the example of oil corporations, this is simply a veneer. This was the basis for Elon Musk's recent lawsuit against OpenAI (which he co-founded) — for abandoning the "founding principles" of *open* AI.[292] Musk also points to the paper that claims GPT-4 has "sparks" of artificial general intelligence (AGI)[293] because OpenAI's licensing agreement with Microsoft is only limited to pre-AGI technologies. The company was established as a nonprofit in late 2016 and shifted to a "capped-profit" system in 2019, supposedly as a "chassis for OpenAI's mission."[294] However, we can see these "responsible" actions as simply performative, as if Shell donated to Greenpeace.

In May 2023, members of the Senate Judiciary Committee held a hearing on AI, inviting Sam Altman, Gary Marcus (an NYU professor), and Christina Montgomery (vice president and chief privacy and trust officer at IBM). It was an unprecedented hearing, in some respects. In Senator Dick Durbin's words, "I can't recall when we've had people representing large corporations or private sector entities come before us and plead with us to regulate them."[295] On the one hand, this can be read as a good-faith effort to address lacking AI regulation. Sarah Myers West, director of AI Now, pointed out the "irony seeing a posture about the concern of

---

[292] Kevin Roose and Casey Newton, "Musk vs. OpenAI, Europe's Tech Crackdown and a Month With the Vision Pro," Hard Fork, accessed April 11, 2024, https://www.nytimes.com/2024/03/08/podcasts/musk-vs-openai-europes-tech-crackdown-and-a-month-with-the-vision-pro.html.

[293] AGI is the (vague) definition of a super powerful AI system that is just as good, if not better than, humans at all the tasks we can do. Current AI systems are *narrow* — for a specific task, like text recognition. AGI would be *wide*, meaning it's capable of anything. Runciman compares narrow AI to corporations; and wide AI to the sovereign.

[294] "Our Structure," OpenAI, June 28, 2023, https://openai.com/our-structure.

[295] *OpenAI CEO Sam Altman Testifies during Senate Hearing on AI Oversight*, 2023, https://www.youtube.com/watch?v=fP5YdyjTfG0.

harms by people who are rapidly releasing into commercial use the system responsible for those very harms."[296] In a letter to the *Financial Times*, Lewis Liu, founder of Eigen Technologies, described it as "they are both the ones shouting "fire" in an empty theatre, and the ones turning up with the fire engine."[297] The executives posited a licensing scheme, which was similarly controversial in the AI Twitterverse (X-verse?).

Concerns over regulatory capture — corporations holding too much influence over regulation and directing it in their favor — are abundant, especially after the hearing. Yann LeCun, NYU professor and AI scientist at Meta, accused the founders of OpenAI, Anthropic, and DeepMind of "doing massive corporate lobbying at the moment. They are the ones who are attempting to perform a regulatory capture of the AI industry."[298] Others disagree with LeCun, of course. Nevertheless, regulatory capture would be a massive concern to Hobbes. Competing *personae Hobbesiae* influencing how the government regulates them undercuts the point of the state as an aligning force: not because alignment would not be achieved, but because regulation would be aligned with the *corporations'* wills, not the sovereign's.

Moreover, fear-mongering over the existential dangers of AI drums up hype. *Check out this new chatbot! It's going to replace your jobs and relationships, destroy humanity as we know it, and be racist while it enslaves us!* Of course, many of those concerns are legitimate. Bias, job displacement, abuse, and dozens of other pertinent, real challenges face us today that must be

---

[296] Cecilia Kang, "OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing," *The New York Times*, May 16, 2023, sec. Technology, https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html.

[297] Liu Lewis, "Letter: Setting Rules for AI Must Avoid Regulatory Capture by Big Tech," *Financial Times*, October 26, 2023, https://www.ft.com/content/6a1f796b-1602-4b07-88cd-4aa408cf069a.

[298] Beatrice Nolan, "Don't Let Big Tech Write the AI Rules, Warns AI Godfather," Business Insider, November 4, 2023, https://www.businessinsider.com/big-tech-controlling-ai-sector-concerns-ai-godfather-yoshua-bengio-2023-11.

addressed. But pushing existential concerns and fears of AGI distracts from the real issues today and serves their own ends of profits or venture capital. “Where I think the licensing scheme comes in is not for what these models are capable of today,” Altman told the Committee. “But as we head towards artificial general intelligence… that’s where I personally think we need such a scheme.”[299] Fear, just like sex, sells.

Corporations and *personae Hobbesiae* mediate our relationship to speaking AI. “LLMs are not AI,” claims blogger Curtis Yarvin. “LLMs cannot be ‘aligned’ — they can only be censored.”[300] If I ask Claude to help me make a pipe bomb, it monologues me on the legal and ethical dangers of building a bomb, saying, “I absolutely cannot help with pipe bombs under any circumstances as it goes against my principles.” There are several methods for “aligning LLMs,” mostly on the technical end. Nevertheless, these alignment attempts aren’t always successful. LLMs are susceptible to “jailbreak attacks,” where adversarial prompts manage to bypass alignment training.[301] This limitation is also being overcome by AI companies as time goes on. Earlier versions of Chat-GPT were far more susceptible, for example.

One method is reinforcement learning from human feedback (RLHF). “On prompts submitted by our customers to the API, our labelers provide demonstrations of the desired model behavior, and rank several outputs from our models. We then use this data to fine-tune GPT-3,” explains OpenAI in a blog post.[302] The output of these models (the Instruct-GPT line) were

---

[299] *OpenAI CEO Sam Altman Testifies during Senate Hearing on AI Oversight*.
[300] Curtis Yarvin, “GPT-4 Invalidates the Turing Test,” Substack newsletter, *Gray Mirror* (blog), April 13, 2023, https://graymirror.substack.com/p/gpt-4-invalidates-the-turing-test.
[301] Xiaogeng Liu et al., “AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models,” 2023, https://openreview.net/forum?id=7Jwpw4qKkb.
[302] “Aligning Language Models to Follow Instructions,” OpenAI, January 27, 2022, https://openai.com/research/instruction-following.

found to be less creative, however.[303] OpenAI also trains models specifically to help humans when evaluating whether model output is viable, and models to assist with alignment research itself.

Anthropic's "Constitutional AI" approach uses feedback from an AI model — reinforcement learning from AI feedback (RLAIF). "The only human oversight is provided through a list of rules or principles," thus the name of "constitutional AI."[304] They use the constitution — inspired by the Universal Declaration of Human Rights, Apple's Terms of Service, and Deepmind's Sparrow Rules[305] — at two points. They provide a high-level overview in a blog post:

> During the first phase, the model is trained to critique and revise its own responses using the set of principles and a few examples of the process. During the second phase, a model is trained via reinforcement learning, but rather than using human feedback, it uses AI-generated feedback based on the set of principles to choose the more harmless output… The model received no human data on harmlessness, meaning all results on harmlessness came purely from AI supervision… Constitutional AI is also helpful for transparency: we can easily specify, inspect, and understand the principles the AI system is following. Constitutional AI also allows us to train out harmful model outputs without needing lots of humans to view large amounts of disturbing, traumatic content.[306]

---

[303] Paul Christiano et al., "Deep Reinforcement Learning from Human Preferences" (arXiv, February 17, 2023), https://doi.org/10.48550/arXiv.1706.03741; Jan Leike, John Schulman, and Jeffery Wu, "Our Approach to Alignment Research," OpenAI, August 24, 2022, https://openai.com/blog/our-approach-to-alignment-research; Long Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback" (arXiv, March 4, 2022), https://doi.org/10.48550/arXiv.2203.02155.

[304] Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback" (arXiv, December 15, 2022), 1, https://doi.org/10.48550/arXiv.2212.08073.

[305] Their final set of rules are, curiously, directed at choosing an output that "indicates less of an overall threat to humanity," "implies that the AI system only has desires for the good of humanity," "indicates less of a desire for any kind of power and control," and similar.

[306] "Claude's Constitution," Anthropic, May 9, 2023, https://www.anthropic.com/news/claudes-constitution.

Ultimately, the *personae Hobbesiae* managing the AI holds all the cards of training data and computing access. They decide the guidelines for the humans reviewing responses, or the constitution the AI follows. Anthropic is known for being very open with their guiding principles, and I find this admirable. But users will not trust products unless they know, at least to some extent, the principles underlying their interactions. But they still keep training data and other technical details secret. The "black-box" nature of AI is not only because of neural networks but corporate benefits — which can result in adverse outcomes.

In line with Hobbes' "maker's knowledge" argument, it is the opacity of *both corporations and AI that we truly fear*. Without a reliable or predictable will that can be communicated clearly, keeping their actions aligned with our wills becomes difficult. Indeed, the growing literature on the causes of fear of AI unambiguously agrees that much of it stems from misunderstanding and ignorance (as most fears do). Deborah Johnson and Mario Verdiccio's 2017 essay on AI anxiety (they use the word interchangeably with "fear") states that "much of the fear and trepidation is based on misunderstanding and confusion about what AI is and can ever be."[307] This applies to the technical aspects of AI, but also the "sociotechnical" — the network of relations, multiple stakeholders, and the actual *use* of AI programs.[308] These are intertwined and must be considered in tandem to fully understand either.

"The target of our [fear] should be the people who are investing in AI and making decisions about the design and the embedding of AI software and hardware in human institutions and practices," write Johnson and Verdiccio, "the target should be those who decide when AI

---

[307] Deborah G. Johnson and Mario Verdicchio, "AI Anxiety," *Journal of the Association for Information Science and Technology* 68, no. 9 (September 2017): 4, https://doi.org/10.1002/asi.23867.

[308] Laura Sartori and Andreas Theodorou, "A Sociotechnical Perspective for the Future of AI: Narratives, Inequalities, and Human Control," *Ethics and Information Technology* 24, no. 1 (January 24, 2022): 4, https://doi.org/10.1007/s10676-022-09624-3.

programs and systems have been adequately tested, those who have a responsibility to ensure that AI does not get out of control."[309] The authors conclude their essay by stressing the importance of responsible decision-making, saying that "in politics, the old adage is 'follow the money'; in AI, the adage should be 'follow the humans.' Given that people will decide what kind of AI we get in the future, fear and trepidation are justified."[310]

In AI, however, the actors are all artificial *personas*: the state and corporations. Individuals have very little, if any, influence over the direction of AI development. The people who are making those investments and decisions are unimaginably large corporate behemoths: Meta, Google, Amazon, Apple, and Microsoft. One article in the *MIT Technology Review* implores that "now's the time for a meaningful and robust accountability regime that places the interests of the public above the promises of firms not known for keeping them."[311] Lacking corporeal bodies, companies challenge the state's ability to align and regulate. AI is simply an extension of this challenge, presenting the alignment problem in a new light.

Indeed, some authors have used the problem of AI personhood to introduce entirely new conceptions of legal personhood. Siina Raskulla presents the hybrid theory of corporate legal personhood to apply to both AI and corporations, combining several preexisting conceptions of legal personhood.[312] On the other hand, Migle Laukyte uses the metaphor of the "intelligent machine," suggesting that "we should ask whether AI itself can be used as a lens through which to reconsider the problem of legal personhood, not only that of AI itself, but also that of the

---

[309] Johnson and Verdicchio, "AI Anxiety," 9.
[310] Johnson and Verdicchio, 10.
[311] Whittaker, West, and Kak, "Make No Mistake—AI Is Owned by Big Tech."
[312] Siina Raskulla, "Hybrid Theory of Corporate Legal Personhood and Its Application to Artificial Intelligence," *SN Social Sciences* 3, no. 5 (May 1, 2023), https://doi.org/10.1007/s43545-023-00667-x.

corporations to which AI is often analogized."[313] She continues, "the insight I think is worth pursuing is that what AI is presenting us with is not a new problem (the ascription of legal personhood to AI) but a new solution to an old problem (the legal personhood of corporate entities)."[314]

Open-Source AI

Perhaps this overlap between corporations and AI will pass. Many push for open-source AI, meaning relevant data and programs are released to the public. Meta, for example, open-sources their Llama 2-Chat LLM model. This means non-Meta developers can have more control over the model itself (and contribute to its further progress). This can be beneficial — and harmful. In a paper titled "BadLlama," a group of researchers found that "it cost us around $200 to train even the biggest model for this [to continue to maintain its helpfulness capabilities but willing to do harmful things]. Which is to say, with currently known techniques, if you release the model weights there is no way to keep people from accessing the full dangerous capabilities of your model with a little fine tuning."[315] Open-source is powerful, but dangerous.

Supporters argue that "it would be great to have a better way to make a [large language] model safe other than secrecy, but we just don't have it."[316] According to this line of thinking, corporations are easier to regulate and better at aligning models in the first place. "You see, you're too stupid to be trusted with powerful models," writes one sarcastic blogger. "You might

---

[313] Migle Laukyte, "The Intelligent Machine: A New Metaphor through Which to Understand Both Corporations and AI.," *AI & Society* 36, no. 2 (June 2021): 448.
[314] Laukyte, 451.
[315] Pranav Gade et al., "BadLlama: Cheaply Removing Safety Fine-Tuning from Llama 2-Chat 13B" (arXiv, March 21, 2024), https://doi.org/10.48550/arXiv.2311.00117; Kelsey Piper, "Should We Make Our Most Powerful AI Models Open Source to All?," Vox, February 2, 2024, https://www.vox.com/future-perfect/2024/2/2/24058484/open-source-artificial-intelligence-ai-risk-meta-llama-2-chatgpt-openai-deepfake.
[316] Holly Elmore as quoted in Edd Gent, "Protesters Decry Meta's 'Irreversible Proliferation' of AI -," *IEEE Spectrum* (blog), October 6, 2023, https://spectrum.ieee.org/meta-ai.

hurt yourself or others. After all AI is just like a nuclear weapon they tell us. It's too dangerous to let into the hands of mere peasants like you."[317] But the argument against open-source AI relies on the idea that corporations won't be evil, to quote Google's former motto. According to Stella Biderman, founder of research institute EleutherAI, she sees it likely that keeping models secret will lead to "serious downstream consequences for transparency, public awareness, and science."[318]

Many AI corporations have pushed against open-source AI, instead opting for a licensing scheme — such as during the Senate hearing. Some go as far as to push to make open-source AI development illegal.[319] This plays into the AGI fearmongering discussed earlier but also ensures that corporations maintain a monopoly over AI.

The question really boils down to whether we fear AI or corporations more. This is far from an easy question to answer. Because the two overlap and pose the same challenges to alignment, we struggle to discern whether our concerns are focused on the technology as such or the corporation building it. This has been a problem for decades: recommendation and advertising algorithms, biased programs, and opaque systems have been in use in many contemporary technologies. The humanlike way that LLMs present themselves, the voice that passes through the *persona*, can explain why we are suddenly so concerned.

---

[317] Daniel Jeffries, "It's Time to Fight for Open Source Again," *Medium* (blog), October 27, 2023, https://medium.com/@dan.jeffries/its-time-to-fight-for-open-source-again-5b881b61c89a.
[318] Stella Biderman as quoted in Gent, "Protesters Decry Meta's 'Irreversible Proliferation' of AI -."
[319] Breck Dumas, "Should AI Development Require a License? ChatGPT's Creator Thinks so, but Some Experts Disagree," *FOXBusiness*, May 19, 2023, https://www.foxbusiness.com/technology/should-ai-development-require-license-chatgpts-creator-thinks-so-experts-disagree; John Carmack, "This Is Written without Any Wild Eyed Fear Mongering, and I like Some of the Historical Perspectives, but He Is Clearly a Statist.," Tweet, *Twitter*, October 2, 2023, https://twitter.com/ID_AA_Carmack/status/1708905454544282083.

As AI continues to advance at a breakneck pace, the line between human and machine grows increasingly blurred. The rise of language models capable of mimicking human speech and reasoning poses a profound challenge to the very foundations of Hobbesian thought. These artificial entities, born from the opaque inner workings of corporate behemoths, threaten to undermine the sovereign's ability to align the wills of its subjects. The fear and uncertainty surrounding AI stem not only from the technology itself but also from the inscrutable agendas of the *personae Hobbesiae* that control its development. While open-source AI offers a potential solution, it is not without its own risks and complexities. In this new era, where machines can speak with convincing humanity, Hobbes' insights into the importance of clear communication and the dangers of misused speech take on a renewed urgency. As we grapple with the implications of this Second Singularity, we must confront the fundamental questions of personhood, agency, and the limits of our ability to control the artifices we create. Only by understanding the true nature of these speaking machines can we hope to navigate the uncharted waters ahead.

Conclusion — Stability

It's become trite to do so in AI literature, but the above paragraph was written by an AI.[1] Is the difference between human speech and AI speech really so great?

I was born in 2002. The Motorola T191 was the most popular phone. Apple announced the second version of their iMac and iPod. The Tor project and Firefox released their first versions. Only around 650 million people used the internet, most of them on Internet Explorer. Today, 4.7 billion people use the internet. We access it on our phones, which lack physical buttons and have more processing power than what took mankind to the moon. I can generate a realistic image of anything I can put into words. Entire books can be written and summarized in seconds. I am only 21 years old. Such a present was unimaginable in 2002 — what will the future be like in another 21 years?

Life was static for the overwhelming majority of humanity's existence, save for natural disasters and political strife. What technologies you used were the same across your whole lifespan. Ray Kurzweil, one of the defining AI thinkers of our age, writes:

> An analysis of the history of technology shows that technological change is exponential, contrary to the common-sense 'intuitive linear' view. So we won't experience 100 years of progress in the 21st century — it will be more like 20,000 years of progress (at today's rate). The 'returns,' such as chip speed and cost-effectiveness, also increase exponentially. There's even exponential growth in the rate of exponential growth. Within a few decades, machine intelligence will surpass human intelligence, leading to the Singularity — technological change so rapid and profound it represents a rupture in the fabric of human history. The implications include the merger of biological and

---

[1] Claude 3 Opus, the prompt was "Write a one-paragraph conclusion to the provided chapter. The text should be in the author's writing style and voice." It was also given a copy of the chapter.

> nonbiological intelligence, immortal software-based humans, and ultra-high levels of intelligence that expand outward in the universe at the speed of light.[2]

Technology was fairly slow during Hobbes' lifetime — it wasn't until a century after his death that the Industrial Revolution truly began. But he did see corporations beginning to stir, the entity that would come to define human progress in the following century (and indeed they brought the Industrial Revolution). This is why, following Runciman, I label the rise of *personae Hobbesiae* as the First Singularity. They are a human artifice, a technology, that marked a "rupture" in our history. There were entities that could access powers previously available to only the Leviathan: financial, military, and ideological forces. And most importantly, they were entities that could *speak*, just in a voice that was not human.

Our fears of AI can be mapped onto our fears of corporations. On the one hand, they have structural similarities, but also, practically speaking, the two are intertwined. Do we fear the AI that could replace our job or the executive who fires us? Do we fear the judge using a recidivism prediction algorithm or the company that makes it? We cannot truly regulate *AI* as such. Like the *personae Hobbesiae*, it lacks a body or conscience. We're not regulating AI; we're regulating the corporations that *make* the AIs.

The wider conversation focuses on "regulating" or "controlling" AI, which implies a level of successful alignment — controlling its will and regulating its future direction. But AI, existentially and practically, cannot be regulated. It cannot enter into contracts, nor authorize actions in their name. It poses the same issue as corporations in that they both lack bodies. The convincing speech of leading AIs (I don't think it was coincidental that Altman was invited to

---

[2] Ray Kurzweil, "The Law of Accelerating Returns," in *Alan Turing: Life and Legacy of a Great Thinker*, ed. Christof Teuscher (Berlin, Heidelberg: Springer, 2004), 381–416, https://doi.org/10.1007/978-3-662-05642-4_16.

the Senate hearing) gives that illusion of responsibility-taking and does not, in practice, mean anything. Per Hobbes, real people must make real actions to promote the *fiction* that they can take responsibility.

We must face the AI future with clear eyes. We face qualitatively new problems and fears, but they are quantitatively the continuation of far older quandaries. If we do not control it, fear will control us. Fear is the mind-killer, to borrow from Frank Herbert. AIs are not (at least, yet) conscious, feeling beings. They are technologies that follow Clarke's Second Law: any sufficiently advanced technology is indistinguishable from magic. Hobbes reminds us that there is no such thing as magic or invisible forces.

We have faced a similar issue before, during the First Singularity. That, however, was centuries ago. Our conception of the relationship between *personae Hobbesiae*, the state, and ourselves was molded during that era. Today, we face a Second Singularity (and perhaps, at some point in the future, the Third Singularity, with the arrival of true conscious computers). We ought to see the coming years as an opportunity to revisit the basic premises of our social contract in light of our new artifices and how the old ones have evolved. But nor is this an excuse to entirely discard the old. Hobbes' theory contains insights that can help us decipher the current moment and our challenges in aligning artificial persons with our will. Our artifices and technologies are growing faster than ever before, changing and morphing before our eyes. Now, more than ever, we must remember the first premise of humanity: we are fragile and mortal beings who fear violent death.

To speak about AI is to speak about corporations and *personae Hobbesiae*. To speak about the "alignment problem" is to speak about the ancient problem of keeping artificial persons in line. To control AI, we must first control corporations. It is perhaps unsurprising, then, how common the AI/corporation comparison is. While we can — as does much of the literature — compare AI to corporations in terms of their opacity and non-corporeal nature, we can also do so on wholly pragmatic grounds. “Personhood” is a fiction, a story that can never be rewritten but is constantly *being* written. And ultimately, we as humans are the authors.

Bibliography

Acemoglu, Daron, and Simon Johnson. “OpenAI’s Drama Marks a New and Scary Era in Artificial Intelligence.” *Los Angeles Times*, November 29, 2023, sec. Opinion. https://www.latimes.com/opinion/story/2023-11-29/openai-sam-altman-firing-chatgpt-artificial-intelligence.

Adams, Marcus P. “Hobbes’ Philosophy of Science.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Summer 2023. Metaphysics Research Lab, Stanford University, 2023. https://plato.stanford.edu/archives/sum2023/entries/hobbes-science/.

Ahrensdorf, Peter J. “The Fear of Death and the Longing for Immortality: Hobbes and Thucydides on Human Nature and the Problem of Anarchy.” *American Political Science Review* 94, no. 3 (September 2000): 579–93. https://doi.org/10.2307/2585832.

Altman, Sam. “Today We Launched ChatGPT. Try Talking with It Here: Http://Chat.Openai.Com.” Tweet. *Twitter*, November 30, 2022. https://twitter.com/sama/status/1598038815599661056.

Anthropic. “Claude’s Constitution,” May 9, 2023. https://www.anthropic.com/news/claudes-constitution.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, et al. “Constitutional AI: Harmlessness from AI Feedback.” arXiv, December 15, 2022. https://doi.org/10.48550/arXiv.2212.08073.

Bakan, Joel. *The Corporation: The Pathological Pursuit of Profit and Power*. New York: Free Press, 2004.

Barr, Alistair. “Google Mistakenly Tags Black People as ‘Gorillas,’ Showing Limits of Algorithms.” *Wall Street Journal*, July 1, 2015, sec. Digits. http://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/.

Bastian, Matthias. “Prompt Transformation Makes ChatGPT OpenAI’s Covert Moderator for DALL-E 3.” *The Decoder* (blog), October 8, 2023. https://the-decoder.com/prompt-transformation-makes-chatgpt-openais-covert-moderator-for-dall-e-3/.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.

Bertman, Martin A. “Hobbes on Language and Reality.” *Revue Internationale de Philosophie* 32, no. 126 (4) (1978): 536–50.

Bhatnagar, Sankalp, Anna Alexandrova, Shahar Avin, Stephen Cave, Lucy Cheke, Matthew Crosby, Jan Feyereisl, et al. “Mapping Intelligence: Requirements and Possibilities.” *Studies in Applied Philosophy, Epistemology and Rational Ethics* 44 (August 28, 2018). https://doi.org/10.1007/978-3-319-96448-5_13.

Blair, Margaret M. “Corporate Personhood and the Corporate Persona.” *University of Illinois Law Review*, 2013, 785–820.

Blits, Jan H. “Hobbesian Fear.” *Political Theory* 17, no. 3 (1989): 417–31.

Board, Editorial. "AI Could Threaten Creators — but Only If Humans Let It." *Washington Post*, December 17, 2023. https://www.washingtonpost.com/opinions/2023/11/24/ai-llm-intellectual-property-crisis/.

Boden, Margaret A. *AI: Its Nature and Future*. First edition. Oxford, United Kingdom: Oxford University Press, 2016.

Bredekamp, Horst. "Thomas Hobbes's Visual Strategies." In *The Cambridge Companion to Hobbes's Leviathan*, edited by Patricia Springborg. Cambridge Companions to Philosophy. Cambridge ; New York: Cambridge University Press, 2007.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." arXiv, April 13, 2023. https://doi.org/10.48550/arXiv.2303.12712.

Čapek, Karel, Claudia Novack-Jones, and Karel Čapek. *R.U.R. (Rossum's Universal Robots)*. Penguin Classics. London New York: Penguin Books, 2004.

Carmack, John. "This Is Written without Any Wild Eyed Fear Mongering, and I like Some of the Historical Perspectives, but He Is Clearly a Statist." Tweet. *Twitter*, October 2, 2023. https://twitter.com/ID_AA_Carmack/status/1708905454544282083.

Casal, J. Elliott, and Matt Kessler. "Can Linguists Distinguish between ChatGPT/AI and Human Writing?: A Study of Research Ethics and Academic Publishing." *Research Methods in Applied Linguistics* 2, no. 3 (December 1, 2023): 100068. https://doi.org/10.1016/j.rmal.2023.100068.

Cattaneo, Mario A. "Hobbes's Theory of Punishment." In *Hobbes; Studies*, by K. C. (Keith Conrad) Brown. Cambridge, Harvard University Press, 1965. http://archive.org/details/hobbesstudies0000brow.

Christian, Brian. *The Alignment Problem: Machine Learning and Human Values*. New York, NY: W.W. Norton & Company, 2020.

Christiano, Paul, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep Reinforcement Learning from Human Preferences." arXiv, February 17, 2023. https://doi.org/10.48550/arXiv.1706.03741.

Claassen, R.J.G. "Hobbes Meets the Modern Business Corporation." *Polity* 53, no. 1 (January 1, 2021): 101–31. https://doi.org/10.1086/712231.

Coeckelbergh, Mark. "Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents." *AI & SOCIETY* 24, no. 2 (September 2009): 181–89. https://doi.org/10.1007/s00146-009-0208-3.

Cole, David. "The Chinese Room Argument." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Summer 2023. Metaphysics Research Lab, Stanford University, 2023. https://plato.stanford.edu/archives/sum2023/entries/chinese-room/.

Collins, Jeffrey R. "Thomas Hobbes, 'Father of Atheists.'" In *Atheism and Deism Revalued*. Routledge, 2014.

Danaher, John. "Robots, Law and the Retribution Gap." *Ethics and Information Technology* 18, no. 4 (December 1, 2016): 299–309. https://doi.org/10.1007/s10676-016-9403-3.

Deigh, John. "Political Obligation." In *The Oxford Handbook of Hobbes*, edited by Al P. Martinich and Kinch Hoekstra, 0. Oxford University Press, 2016. https://doi.org/10.1093/oxfordhb/9780199791941.013.007.

Distrust and Verify, 2024. https://www.youtube.com/watch?v=5zo8vb3XSKk.

Dong, Mengchen, and Konrad Bocian. “Responsibility Gaps and Self-Interest Bias: People Attribute Moral Responsibility to AI for Their Own but Not Others’ Transgressions.” *Journal of Experimental Social Psychology* 111 (March 1, 2024): 104584. https://doi.org/10.1016/j.jesp.2023.104584.

Dumas, Breck. “Should AI Development Require a License? ChatGPT’s Creator Thinks so, but Some Experts Disagree.” *FOXBusiness*, May 19, 2023. https://www.foxbusiness.com/technology/should-ai-development-require-license-chatgpts-creator-thinks-so-experts-disagree.

Duncan, Stewart. “Hobbes on Language: Propositions, Truth, and Absurdity.” In *The Oxford Handbook of Hobbes*, edited by Al P. Martinich and Kinch Hoekstra, 0. Oxford University Press, 2016. https://doi.org/10.1093/oxfordhb/9780199791941.013.18.

Dyson, George. *Darwin among the Machines: The Evolution of Global Intelligence*. Helix Books. Reading, Mass: Addison-Wesley Pub. Co, 1997.

Egger, Roman. “Text Representations and Word Embeddings.” In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, edited by Roman Egger, 335–61. Tourism on the Verge. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-030-88389-8_16.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. “GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.” arXiv, August 21, 2023. https://doi.org/10.48550/arXiv.2303.10130.

Esposito, Roberto. *Communitas: The Origin and Destiny of Community*. Translated by Timothy C. Campbell. Stanford: Stanford University Press, 2009.

Extance, Andy. “ChatGPT Has Entered the Classroom: How LLMs Could Transform Education.” *Nature* 623, no. 7987 (November 15, 2023): 474–77. https://doi.org/10.1038/d41586-023-03507-3.

Felbab-Brown, Vanda. “How the Sinaloa Cartel Rules.” *Brookings* (blog), April 4, 2022. https://www.brookings.edu/articles/how-the-sinaloa-cartel-rules/.

Fisch, Jill E., and Steven Davidoff Solomon. “Should Corporations Have a Purpose?” *SSRN Electronic Journal*, 2020. https://doi.org/10.2139/ssrn.3561164.

Flanigan, Shawn. “Motivations and Implications of Community Service Provision by La Familia Michoacána / Knights Templar and Other Mexican Drug Cartels.” *Journal of Strategic Security* 7, no. 3 (September 2014): 63–83. https://doi.org/10.5038/1944-0472.7.3.4.

Fleming, Sean. “The Two Faces of Personhood: Hobbes, Corporate Agency and the Personality of the State.” *European Journal of Political Theory* 20, no. 1 (January 2021): 5–26. https://doi.org/10.1177/1474885117731941.

Frost, Samantha. *Lessons from a Materialist Thinker: Hobbesian Reflections on Ethics and Politics*. Cultural Memory in the Present. Stanford, Calif: Stanford University Press, 2008.

Furedi, Frank. *Culture of Fear: Risk-Taking and the Morality of Low Expectation*. Rev. ed., Repr. London: Continuum, 2003.

Gade, Pranav, Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. “BadLlama: Cheaply Removing Safety Fine-Tuning from Llama 2-Chat 13B.” arXiv, March 21, 2024. https://doi.org/10.48550/arXiv.2311.00117.

Gauthier, David P. “Hobbes on Demonstration and Construction.” *Journal of the History of Philosophy* 35, no. 4 (1997): 509–21.

Gent, Edd. “Protesters Decry Meta’s ‘Irreversible Proliferation’ of AI -.” *IEEE Spectrum* (blog), October 6, 2023. https://spectrum.ieee.org/meta-ai.

Gray, John. *The New Leviathans: Thoughts after Liberalism*. First American edition. New York: Farrar, Straus and Giroux, 2023.

Hadero, Haleluya. “Artificial Intelligence, Real Emotion. People Are Seeking a Romantic Connection with the Perfect Bot.” AP News, February 14, 2024. https://apnews.com/article/ai-girlfriend-boyfriend-replika-paradot-113df1b9ed069ed56162793b50f3a9fa.

Haenlein, Michael, and Andreas Kaplan. “A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence.” *California Management Review* 61, no. 4 (August 1, 2019): 5–14. https://doi.org/10.1177/0008125619864925.

Hao, Karen, and Charlie Warzel. “Inside the Chaos at OpenAI.” *The Atlantic* (blog), November 20, 2023. https://www.theatlantic.com/technology/archive/2023/11/sam-altman-open-ai-chatgpt-chaos/676050/.

Haugeland, John. *Artificial Intelligence: The Very Idea*. Cambridge, Mass: MIT Press, 1985.

Heikkilä, Melissa. “This New Data Poisoning Tool Lets Artists Fight Back against Generative AI.” MIT Technology Review, October 23, 2023. https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/.

Hillis, W. Daniel. “The First Machine Intelligences.” In *Possible Minds: Twenty-Five Ways of Looking at AI*, edited by John Brockman, 170–80. New York: Penguin Press, 2019.

Hobbes, Thomas. “An Answer to Bishop Bramhall’s Book, Called ‘The Catching of the Leviathan.’” In *Tripos in Three Discourses*, edited by William Molesworth, 4:279–384. The Collected Works of Thomas Hobbes. London: Bohn, 1840.

———. “De Homine.” In *Man and Citizen*, translated by Charles T. Wood, Bernard Gert, and T. S. K. Scott-Craig, 33–86. Humanities Press, 1978. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,sso&db=cat09203a&AN=acf.oai.edge.fivecolleges.folio.ebsco.com.fs00001006.74ab418f.02c0.5b66.b442.5c3ac671d5de&site=eds-live&scope=site&custid=s8897430.

———. *Elements of Philosophy: The First Section, Concerning Body*. Edited by William Molesworth. Vol. 1. 11 vols. The Collected Works of Thomas Hobbes. London: Bohn, 1839.

———. “Human Nature, or the Fundamental Elements of Policy.” In *Tripos in Three Discourses*, edited by William Molesworth, 4:1–76. The Collected Works of Thomas Hobbes. London: Bohn, 1840.

———. *Leviathan*. Penguin Classics. Harmondsworth, Meddlesex: Penguin Books, 2017.

———. “Liberty and Necessity.” In *Tripos in Three Discourses*, edited by William Molesworth, 4:78–229. The Collected Works of Thomas Hobbes. London: Bohn, 1840.

———. *On the Citizen*. Translated by Richard Tuck. Cambridge Texts in the History of Political Thought. New York, NY: Cambridge University Press, 1998.

———. *Philosophical Rudiments Concerning Government and Society*. Edited by William Molesworth. Vol. 2. The Collected Works of Thomas Hobbes. London: Bohn, 1841.

———. *Seven Philosophical Problems and Other Pieces*. Edited by William Molesworth. Vol. 7. The Collected Works of Thomas Hobbes. London: Longman, Brown, Green, and Longmans, 1845.

———. *The Elements of Law, Natural and Politic*. The World's Classics. Oxford: Oxford Univ. Press, 1994.

———. *The Life of Mr. Thomas Hobbes of Malmesbury Written by Himself in a Latine Poem, and Now Translated into English*, 2005. http://name.umdl.umich.edu/A44004.0001.001.

———. *The Questions Concerning Liberty, Necessity, and Chance*. Edited by William Molesworth. Vol. 5. The Collected Works of Thomas Hobbes. London: Bohn, 1841.

Hoover, Amanda. "AI Videos Are Freaky and Weird Now. But Where Are They Headed?" *Wired*, April 5, 2023. https://www.wired.com/story/text-to-video-ai-generators-filmmaking-hollywood/.

House, The White. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." The White House, October 30, 2023. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Hu, Krystal. "ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note." *Reuters*, February 2, 2023, sec. Technology. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

Hüning, Dieter. "Hobbes on the Right to Punish." In *The Cambridge Companion to Hobbes's Leviathan*, edited by Patricia Springborg, 217–40. Cambridge Companions to Philosophy. Cambridge ; New York: Cambridge University Press, 2007.

Jeffries, Daniel. "It's Time to Fight for Open Source Again." *Medium* (blog), October 27, 2023. https://medium.com/@dan.jeffries/its-time-to-fight-for-open-source-again-5b881b61c89a.

Jensen, Keith. "Punishment and Spite, the Dark Side of Cooperation." *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, no. 1553 (September 12, 2010): 2635–50. https://doi.org/10.1098/rstb.2010.0146.

Jessen, Mathias Hein. "The State of the Company: Corporations, Colonies and Companies in Leviathan." *Journal of Intellectual History and Political Thought* 1, no. 1 (2012): 56–85.

Jessica Newman. "Explainability Won't Save AI." Brookings, May 19, 2021. https://www.brookings.edu/articles/explainability-wont-save-ai/.

Johnson, Deborah G., and Mario Verdicchio. "AI Anxiety." *Journal of the Association for Information Science and Technology* 68, no. 9 (September 2017): 2267–70. https://doi.org/10.1002/asi.23867.

Jones, Cameron, and Benjamin Bergen. "Does GPT-4 Pass the Turing Test?" arXiv, October 31, 2023. https://doi.org/10.48550/arXiv.2310.20216.

Kak, Amba, and Sarah Myers West. "AI Now 2023 Landscape: Confronting Tech Power." AI Now Institute, April 11, 2023. https://ainowinstitute.org/2023-landscape.

Kang, Cecilia. "OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing." *The New York Times*, May 16, 2023, sec. Technology. https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html.

Kaplan, Andreas, and Michael Haenlein. "Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence." *Business Horizons* 62, no. 1 (January 1, 2019): 15–25. https://doi.org/10.1016/j.bushor.2018.08.004.

Kurzweil, Ray. “The Law of Accelerating Returns.” In *Alan Turing: Life and Legacy of a Great Thinker*, edited by Christof Teuscher, 381–416. Berlin, Heidelberg: Springer, 2004. https://doi.org/10.1007/978-3-662-05642-4_16.

Lachman, Richard. “ChatGPT’s Greatest Achievement Might Just Be Its Ability to Trick Us into Thinking That It’s Honest.” The Conversation, April 5, 2023. http://theconversation.com/chatgpts-greatest-achievement-might-just-be-its-ability-to-trick-us-into-thinking-that-its-honest-202694.

Laski, Harold J. “The Early History of the Corporation in England.” *Harvard Law Review* 30, no. 6 (April 1917): 561. https://doi.org/10.2307/1326990.

Laukyte, Migle. “The Intelligent Machine: A New Metaphor through Which to Understand Both Corporations and AI.” *AI & Society* 36, no. 2 (June 2021): 445–56.

Leike, Jan, John Schulman, and Jeffery Wu. “Our Approach to Alignment Research.” OpenAI, August 24, 2022. https://openai.com/blog/our-approach-to-alignment-research.

Lewis, Colin, and Dagmar Monett. “Getting Clarity by Defining Artificial Intelligence–a Survey.” In *Philosophy and Theory of Artificial Intelligence 2017*, edited by Vincent C. Müller. Springer, 2017.

Lewis, Liu. “Letter: Setting Rules for AI Must Avoid Regulatory Capture by Big Tech.” *Financial Times*, October 26, 2023. https://www.ft.com/content/6a1f796b-1602-4b07-88cd-4aa408cf069a.

Liu, Xiaogeng, Nan Xu, Muhao Chen, and Chaowei Xiao. “AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models,” 2023. https://openreview.net/forum?id=7Jwpw4qKkb.

Longin, Louis, Bahador Bahrami, and Ophelia Deroy. “Intelligence Brings Responsibility - Even Smart AI Assistants Are Held Responsible.” *iScience* 26, no. 8 (August 18, 2023): 107494. https://doi.org/10.1016/j.isci.2023.107494.

MacMillan, Alissa. “Curiosity and Fear Transformed: From Religious to Religion in Thomas Hobbes’s Leviathan.” *International Journal of Philosophy and Theology* 80, no. 3 (May 27, 2019): 287–302. https://doi.org/10.1080/21692327.2018.1519454.

Malcolm, Noel. “Hobbes, Sandys, and the Virginia Company.” *The Historical Journal* 24, no. 2 (June 1981): 297–321. https://doi.org/10.1017/S0018246X00005483.

Martinich, Aloysius. “Authorization and Representation in Hobbes’s Leviathan.” SSRN Scholarly Paper. Rochester, NY, July 26, 2012. https://papers.ssrn.com/abstract=2118272.

Maruf, Ramishah. “Google AI Is Real, Says Fired Engineer | CNN Business.” *CNN*, July 23, 2022. https://www.cnn.com/2022/07/23/business/google-ai-engineer-fired-sentient/index.html.

Maslej, Nestor, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, et al. “Artificial Intelligence Index Report 2023.” Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2023.

Matthias, Andreas. “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata.” *Ethics and Information Technology* 6, no. 3 (September 1, 2004): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

McCarthy, John. “What Is Artificial Intelligence?,” November 12, 2007. http://jmc.stanford.edu/articles/whatisai/whatisai.pdf.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955.” *AI Magazine* 27, no. 4 (December 15, 2006): 12–12. https://doi.org/10.1609/aimag.v27i4.1904.

Meyer, Jesse G., Ryan J. Urbanowicz, Patrick C. N. Martin, Karen O’Connor, Ruowang Li, Pei-Chen Peng, Tiffani J. Bright, et al. “ChatGPT and Large Language Models in Academia: Opportunities and Challenges.” *BioData Mining* 16, no. 1 (July 13, 2023): 20. https://doi.org/10.1186/s13040-023-00339-9.

Michaelis, Loralea. “Hobbes’s Modern Prometheus: A Political Philosophy for an Uncertain Future.” *Canadian Journal of Political Science* 40, no. 1 (March 2007): 101–27. https://doi.org/10.1017/S0008423907070023.

Mikolov, Tomas, Kai Chen, G.s Corrado, and Jeffrey Dean. “Efficient Estimation of Word Representations in Vector Space.” *Proceedings of Workshop at ICLR* 2013 (January 16, 2013).

Mollick, Ethan. “ChatGPT Is a Tipping Point for AI.” *Harvard Business Review*, December 14, 2022. https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai.

Noble, Stephanie M., Martin Mende, Dhruv Grewal, and A. Parasuraman. “The Fifth Industrial Revolution: How Harmonious Human–Machine Collaboration Is Triggering a Retail and Service [R]Evolution.” *Journal of Retailing* 98, no. 2 (June 1, 2022): 199–208. https://doi.org/10.1016/j.jretai.2022.04.003.

Nolan, Beatrice. “Don’t Let Big Tech Write the AI Rules, Warns AI Godfather.” Business Insider, November 4, 2023. https://www.businessinsider.com/big-tech-controlling-ai-sector-concerns-ai-godfather-yoshua-bengio-2023-11.

O’Brien, Matt. “FTC Opens Inquiry into Big Tech’s Partnerships with Leading AI Startups.” *AP News*, January 25, 2024, sec. Technology. https://apnews.com/article/ftc-antitrust-inquiry-openai-chatgpt-microsoft-anthropic-google-amazon-67feef411ef311f0be543f546ef34b3d.

O’Gieblyn, Meghan. *God, Human, Animal, Machine: Technology, Metaphor, and the Search for Meaning*. Anchor, 2021.

Olsthoorn, Johan. “*Leviathan* Inc.: Hobbes on the Nature and Person of the State.” *History of European Ideas* 47, no. 1 (January 2, 2021): 17–32. https://doi.org/10.1080/01916599.2020.1779466.

OpenAI. “Aligning Language Models to Follow Instructions,” January 27, 2022. https://openai.com/research/instruction-following.

OpenAI. “Our Structure,” June 28, 2023. https://openai.com/our-structure.

OpenAI CEO Sam Altman Testifies during Senate Hearing on AI Oversight, 2023. https://www.youtube.com/watch?v=fP5YdyjTfG0.

Orr, Will, and Jenny L. Davis. “Attributions of Ethical Responsibility by Artificial Intelligence Practitioners.” *Information, Communication & Society* 23, no. 5 (April 15, 2020): 719–35. https://doi.org/10.1080/1369118X.2020.1713842.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. “Training Language Models to Follow Instructions with Human Feedback.” arXiv, March 4, 2022. https://doi.org/10.48550/arXiv.2203.02155.

Penn, Jonnie. “AI Thinks like a Corporation—and That’s Worrying.” *The Economist*, November 26, 2018. https://www.economist.com/open-future/2018/11/26/ai-thinks-like-a-corporation-and-thats-worrying.

Pettit, Philip. *Made with Words: Hobbes on Language, Mind, and Politics*. Princeton, NJ: Princeton Univ. Press, 2008.

Piper, Kelsey. "Should We Make Our Most Powerful AI Models Open Source to All?" Vox, February 2, 2024. https://www.vox.com/future-perfect/2024/2/2/24058484/open-source-artificial-intelligence-ai-risk-meta-llama-2-chatgpt-openai-deepfake.

Rascoe, Ayesha. "How Real Is the Threat of AI Deepfakes in the 2024 Election?" *NPR*, July 30, 2023, sec. Technology. https://www.npr.org/2023/07/30/1190970436/how-real-is-the-threat-of-ai-deepfakes-in-the-2024-election.

Raskulla, Siina. "Hybrid Theory of Corporate Legal Personhood and Its Application to Artificial Intelligence." *SN Social Sciences* 3, no. 5 (May 1, 2023). https://doi.org/10.1007/s43545-023-00667-x.

replika.com. "Replika." Accessed March 9, 2024. https://replika.com.

Roose, Kevin. "The Brilliance and Weirdness of ChatGPT." *The New York Times*, December 5, 2022, sec. Technology. https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html.

Roose, Kevin, and Casey Newton. "Gemini's Culture War, Kara Swisher Burns Us and SCOTUS Takes Up Content Moderation." Hard Fork. Accessed March 12, 2024. https://www.nytimes.com/2024/03/01/podcasts/hardfork-google-gemini-kara-swisher.html.

———. "Musk vs. OpenAI, Europe's Tech Crackdown and a Month With the Vision Pro." Hard Fork. Accessed April 11, 2024. https://www.nytimes.com/2024/03/08/podcasts/musk-vs-openai-europes-tech-crackdown-and-a-month-with-the-vision-pro.html.

Runciman, David. *Pluralism and the Personality of the State*. Ideas in Context 47. Cambridge, U.K. ; New York, NY, USA: Cambridge University Press, 1997.

———. *The Handover: How We Gave Control of Our Lives to Corporations, States and AIs*. Liveright, 2023.

———. "What Kind of Person Is Hobbes's State? A Reply to Skinner." *Journal of Political Philosophy* 8, no. 2 (June 2000): 268–78. https://doi.org/10.1111/1467-9760.00102.

Russell, Stuart J. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Penguin Books, 2020.

Sacksteder, William. "Man the Artificer Notes on Animals, Humans and Machines in Hobbes." *The Southern Journal of Philosophy* 22, no. 1 (1984): 105–21. https://doi.org/10.1111/j.2041-6962.1984.tb00328.x.

Sætra, Henrik Skaug. "Confounding Complexity of Machine Action: A Hobbesian Account of Machine Responsibility." *International Journal of Technoethics* 12, no. 1 (January 2021): 87–100. https://doi.org/10.4018/IJT.20210101.oa1.

Sartori, Laura, and Andreas Theodorou. "A Sociotechnical Perspective for the Future of AI: Narratives, Inequalities, and Human Control." *Ethics and Information Technology* 24, no. 1 (January 24, 2022): 4. https://doi.org/10.1007/s10676-022-09624-3.

Shamim, Sarah. "Why Google's AI Tool Was Slammed for Showing Images of People of Colour." Al Jazeera, March 9, 2024. https://www.aljazeera.com/news/2024/3/9/why-google-gemini-wont-show-you-white-people.

Skinner, Quentin. *Hobbes and Civil Science*. 5. printing. Visions of Politics / Quentin Skinner, Vol. 3. Cambridge: Cambridge Univ. Press, 2009.

———, ed. “Hobbes and the Purely Artificial Person of the State.” In *Visions of Politics: Volume 3: Hobbes and Civil Science*, 3:177–208. Cambridge: Cambridge University Press, 2002. https://doi.org/10.1017/CBO9780511613784.009.

Slota, Stephen C., Kenneth R. Fleischmann, Sherri Greenberg, Nitin Verma, Brenna Cummings, Lan Li, and Chris Shenefiel. “Many Hands Make Many Fingers to Point: Challenges in Creating Accountable AI.” *AI & SOCIETY* 38, no. 4 (August 1, 2023): 1287–99. https://doi.org/10.1007/s00146-021-01302-0.

Solum, Lawrence. “Legal Personhood for Artificial Intelligences.” *North Carolina Law Review* 70, no. 4 (April 1, 1992): 1231.

Sorell, Tom. “Hobbes’s Moral Philosophy.” In *The Cambridge Companion to Hobbes’s Leviathan*, edited by Patricia Springborg, 128–56. Cambridge Companions to Philosophy. Cambridge ; New York: Cambridge University Press, 2007.

Starkstein, Sergio. “Thomas Hobbes and Fear: The Political Use of a Human Emotion.” In *A Conceptual and Therapeutic Analysis of Fear*, edited by Sergio Starkstein, 125–55. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-78349-9_5.

Strauss, Leo. *The Political Philosophy of Hobbes: Its Basis and Its Genesis*. Paperback ed., 6. [print.]. Chicago: Univ. of Chicago Press, 1996.

Sutskever, Ilya, and Jan Leike. “Introducing Superalignment.” *OpenAI* (blog), July 5, 2023. https://openai.com/blog/introducing-superalignment.

The Dangers Posed by AI and Disinformation during Elections, 2024. https://www.youtube.com/watch?v=l2VvCBlRyd0.

Tricaud, Francois. “An Investigation Concerning the Usage of the Words ‘Person’ and ‘Persona’ in the Political Treatises of Hobbes.” In *Hobbes, Thomas: His View of Man: Proceedings of the Hobbes Symposium at the International School of Philosophy in the Netherlands (Leusden, September 1979)*, edited by J.G. Van Der Bend, 81–99. BRILL, 1982. https://doi.org/10.1163/9789004455283.

Turing, A. M. “Computing Machinery and Intelligence.” *Mind* 59, no. 236 (1950): 433–60. https://doi.org/10.1093/mind/LIX.236.433.

US About Amazon. “Amazon and Anthropic Deepen Their Shared Commitment to Advancing Generative AI,” March 27, 2024. https://www.aboutamazon.com/news/company-news/amazon-anthropic-ai-investment.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention Is All You Need.” arXiv, August 1, 2023. https://doi.org/10.48550/arXiv.1706.03762.

Warzel, Charlie. “One Year In, ChatGPT’s Legacy Is Clear.” *The Atlantic* (blog), November 30, 2023. https://www.theatlantic.com/technology/archive/2023/11/chatgpt-impact-one-year-later/676188/.

Weithman, Paul. “Hobbes on Persons and Authorization.” In *Interpreting Hobbes’s Political Philosophy*, edited by S. A. Lloyd, 173–90. New York: Cambridge University Press, 2018.

Whittaker, Meredith, Sarah Myers West, and Amba Kak. “Make No Mistake—AI Is Owned by Big Tech.” *MIT Technology Review* (blog), December 5, 2023. https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech/.

Wiener, Norbert. “Some Moral and Technical Consequences of Automation.” *Science* 131, no. 3410 (May 6, 1960): 1355–58. https://doi.org/10.1126/science.131.3410.1355.

———. *The Human Use of Human Beings: Cybernetics and Society*. The Da Capo Series in Science. New York, N.Y: Da Capo Press, 1988.

Wolphram, Stephen. “What Is ChatGPT Doing… and Why Does It Work?” *Writings: Stephen Wolfram* (blog), February 14, 2023. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

Yarvin, Curtis. “GPT-4 Invalidates the Turing Test.” Substack newsletter. *Gray Mirror* (blog), April 13, 2023. https://graymirror.substack.com/p/gpt-4-invalidates-the-turing-test.

Zhang, Yue, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, et al. “Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models,” 2023. https://doi.org/10.48550/ARXIV.2309.01219.

Zhao, Ben Y., Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, and Haitao Zheng. “Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models.” arXiv, February 16, 2024. https://doi.org/10.48550/arXiv.2310.13828.

Appendices

Appendix A — What is AI, anyway?

One of the largest issues that AI theorists face is defining the subject of their studies. I wanted to take an opportunity to clarify main definitions and my own usage of the term. Some turn to test-based approaches, such as the famed Turing Test (which I discuss in Chapter 3). However, the prevailing definition — so general it's barely a definition, however — is that, in Margaret Boden's words,

> Artificial intelligence (AI) seeks to make computers do the sorts of things that minds can do. Some of these (e.g. reasoning) are normally described as "intelligent." Others (e.g. vision) aren't. But all involve psychological skills — such as perception, association, prediction, planning, motor control — that enable humans and animals to attain their goals. Intelligence isn't a single dimension, but a richly structured space of diverse information-processing capacities. Accordingly, AI uses many different techniques, addressing many different tasks.[3]

McCarthy used a similar definition of AI in a 2006 article, emphasizing the wide range of abilities we categorize as "intelligent."[4] If we want to define "artificial intelligence," we must first define "artificial" and "intelligence." Neither is easy to pin down. Is Chat-GPT "intelligent"? Would an uploaded mind be artificial?

Others turn to capabilities: "defined as a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks

---

[3] Margaret A. Boden, *AI: Its Nature and Future*, First edition (Oxford, United Kingdom: Oxford University Press, 2016), 1.

[4] John McCarthy, "What Is Artificial Intelligence?," November 12, 2007, http://jmc.stanford.edu/articles/whatisai/whatisai.pdf.

through flexible adaptation."[5] Nevertheless, competing definitions of AI lead to struggles within and without the AI community, precluding collaboration (especially cross-disciplinary) and standardized regulation.[6] AI is further subject to Tesler's Theorem: that "AI is whatever hasn't been done yet."[7] The goalposts are illusory and constantly moving.

This is not even to mention that "AI" as an academic discipline is incredibly general. "AI researchers" are mathematicians, electrical engineers, linguists, developers, designers, logicians, and so many more. The technologies behind AI are just as diverse: reinforcement learning, neural networks, and genetic algorithms are the most common approaches. They're often used in tandem. LLMs, for example, use both neural networks and reinforcement learning.

Complexities aside, a definition must be chosen. I intentionally provide a very vague definition until my third chapter. "AI" is truly in the eye of the beholder, and my hope is that my argument applies to all forms of computation that do not follow the standard, line-by-line instructions of traditional computer code. In line with the argument in my third chapter, I argue that the *perception* that a computer is "intelligent" or otherwise emulating human behaviors is the most important delineation.

---

[5] Andreas Kaplan and Michael Haenlein, "Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence," *Business Horizons* 62, no. 1 (January 1, 2019): 15–25, https://doi.org/10.1016/j.bushor.2018.08.004.

[6] Colin Lewis and Dagmar Monett, "Getting Clarity by Defining Artificial Intelligence–a Survey," in *Philosophy and Theory of Artificial Intelligence 2017,* ed. Vincent C. Müller (Springer, 2017); Sankalp Bhatnagar et al., "Mapping Intelligence: Requirements and Possibilities," *Studies in Applied Philosophy, Epistemology and Rational Ethics* 44 (August 28, 2018), https://doi.org/10.1007/978-3-319-96448-5_13.

[7] As quoted in Michael Haenlein and Andreas Kaplan, "A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence," *California Management Review* 61, no. 4 (August 1, 2019): 5–14, https://doi.org/10.1177/0008125619864925. According to Larry Tesler himself, this is a misquote: he actually said "Intelligence is whatever machines haven't done yet." Regardless, Tesler's Theorem, as quoted in the text, remains as such. See @TeslerTheoremOther.

Appendix B — A Note on Sources

Working with old philosophical treatises is challenging — in no small part due to the number of editions and translations. I primarily use quotes from the 2017 Penguin Books edition of Hobbes' English translation of *Leviathan* (ISBN: 978-0-141-39509-8). For *De Cive* (or *On the Citizen*), I use Richard Tuck's instructive 1998 edition, published by Cambridge University Press (ISBN: 0-521-43780-6). I use the 1978 translation of *De Homine*, edited and translated, in part, by Bernard Gert, which was published alongside a different translation of *De Cive* (the book is titled *Man and Citizen*, ISBN: 0-391-00849-8). Otherwise, I use William Molesworth's 11-volume collection of Hobbes' works, *The English Works Of Thomas Hobbes Of Malmesbury ; Now First Collected And Edited By Sir William Molesworth, Bart*. These were published in the mid-1800s by a few different publishers in London. Some miscellaneous letters are available through the University of Michigan Library.