Reclaiming Human Agency in the Age of Artificial Intelligence

AI Research Group
for the
Centre for Digital Culture
of the
Dicastery for Culture and Education of the Holy See

Edited by
Paul Scherz and Brian Patrick Green

PICKWICK *Publications* • Eugene, Oregon

RECLAIMING HUMAN AGENCY
IN THE AGE OF ARTIFICIAL INTELLIGENCE

Pickwick Publications
An Imprint of Wipf and Stock Publishers
199 W. 8th Ave., Suite 3
Eugene, OR 97401

www.wipfandstock.com

SOFTCOVER ISBN: 979-8-3852-7571-7
HARDCOVER ISBN: 979-8-3852-7572-4
EBOOK ISBN: 979-8-3852-7573-1

Cover image generated by Google Gemini.

Dedication

In memory of Pope Francis

Theological Investigations *of* Artificial Intelligence Book Series

Society has crossed the threshold of massive upheavals due to the advancement and proliferation of artificial intelligence and other technologies. The Theological Investigations of Artificial Intelligence book series will offer reflections on this technological revolution and its consequences from the tradition of Catholic social, economic, and ethical thought.

The series will be published in collaboration between the AI Research Group for the Centre for Digital Culture, which is part of the Holy See's Dicastery for Culture and Education, and the *Journal of Moral Theology*. It advances the *Journal of Moral Theology*'s mission of fostering scholarship deeply rooted in traditions of inquiry about the moral life, engaged with contemporary issues, and exploring the interface of Catholic moral theology, philosophy, economics, political philosophy, psychology, and more.

Online versions of the volumes in the Theological Investigations of Artificial Intelligence series are available for free download at jmt.scholasticahq.com. Paper copies may be purchased from Wipf & Stock. This dual approach reflects the *Journal of Moral Theology*'s commitment to the common good, as it makes the scholarship of Catholic theological ethicists broadly available, especially across borders. Additionally, you can find the series listed on the website of the Dicastery for Culture and Education at: www.dce.va.

Series Titles

Encountering Artificial Intelligence: Ethical and Anthropological Investigations, edited by Matthew J. Gaudet, Noreen Herzfeld, Paul Scherz, and Jordan J. Wales (2024)

Reclaiming Human Agency in the Age of Artificial Intelligence, edited by Paul Scherz and Brian Patrick Green (2025)

The AI Research Group is a group of North American theologians, philosophers, and ethicists who have come together at the invitation of the Vatican Centre for Digital Culture, under the auspices of the Dicastery for Culture and Education of the Holy See, to discuss the current and future issues that the continued development of artificial intelligence poses for life and society as we know it. This book is the result of the collaborative efforts of these scholars from 2024 to 2025. The lead authors for this volume were Paul Scherz and Brian Patrick Green, and the contributing authors were Nathan Colaner, Jeremiah Coogan, Mariele Courtois, Heather Foucault-Camm, Matthew J. Gaudet, Noreen Herzfeld, Cory Andrew Labrecque, Anselm Ramelow, OP, Margarita Vega, Andrea Vicini, SJ, and Joseph Vukov.

Table of Contents

Acknowledgments

We would like to express our thanks to all the people and institutions that have helped us in completing this book. The work was inspired by Pope Francis's commitment to addressing the pressing ethical challenges of AI. The Dicastery for Culture and Education, under the leadership of Cardinal José Tolentino de Mendonça and its Secretary Bishop Paul Tighe, made this work possible, and we are very grateful for their support. We are also very grateful for the continuous hard work and support of Angel Gonzalez-Ferrer, director of the Centre for Digital Culture at the Dicastery for Culture and Education, who has been a constant source of inspiration, organization, and encouragement throughout our writing. We would like to thank Fundación Telefónica, Fundación "LaCaixa," and Fundación ProFuturo for their support of the work of the Centre for Digital Culture over these years, which has been crucial to the success of this research project. We would also like to thank our colleges and universities, which provided the time and support necessary for us to complete this work.

The final structure for the book was developed at a plenary meeting of the group in July 2025 at the University of Notre Dame. That conference was organized by the de Nicola Center for Ethics and Culture. We are grateful for the support of Justin Petrisek, Brooke Tranten, Danny O'Callaghan, Margaret McManaway, and Jennifer Martin for the incredible amount of work they put in making sure that the conference went off without a problem. It received the financial support of a Henkels Large Conference Grant from the Franco Family Institute for Liberal Arts and the Public Good, College of Arts and Letters, University of Notre Dame, for which we are thankful. The McGrath Institute for Church Life provided further logistical support. At the conference, the group received extensive comments on the manuscript from Megan Levis Scheirer, Fr. Javier Prades, Walter Scheirer, and Thomas Stapleford, which shaped the

final manuscript. They were very generous in giving their time to make the work better. The education section of the AI Research Group also provided important comments at the event, so thanks to Catherine Moon, Luis Vera, Ann Skeet, Jason Heron, Kevin Gary, Maria Morrow, John Slattery, Taylor Nutter, and Warren von Eschenbach. Thanks to all of you for your support.

Others have made important contributions at different points in the writing process. Matthew Crawford, Yvonne Masakowski, Brett Kagan, and Jordan Amadio joined earlier Zoom meetings by the group to provide their insights on particular topics. Nicholas Ramirez, Abby Tucker, and Joshua Peck provided research assistance during the meetings and during the drafting of the manuscript. The Notre Dame–IBM Tech Ethics Lab, through the assistance of Megan McDermott and Meghan Sullivan, provided important financial support for the editing and publication of the book. The Markkula Center for Applied Ethics at Santa Clara University has provided continuing support for this work, and the generosity of Dan and Charmaine Warmenhoven has been invaluable for making this project possible, for which we are very thankful. We would also like to thank Catherine Osborne for her editorial assistance, as well as M. Therese Lysaught and the staff at the *Journal of Moral Theology* for their assistance in getting the book through the publication process. It is a much better book thanks to all of your contributions.

The Authors

Preface

In the two years that have passed since the publication of *Encountering Artificial Intelligence: Ethical and Anthropological Investigations,* the AI Research Group, established by the Centre for Digital Culture of the Dicastery for Culture and Education, has continued its work of reflection on the likely social and cultural impact of AI. The extraordinarily generous and fruitful contribution of the original researchers and scholars has, if anything, been intensified in the intervening period, which has seen the addition of further participants to the group and the emergence of an even more ambitious program of work. I am very pleased to welcome the latest fruit of the group's work: *Reclaiming Human Agency in the Age of AI.*

The text, in continuity with the group's previous publication and following *Antiqua et Nova,*[1] highlights the difference between the prevailing AI definition of intelligence as rational action and problem-solving (often influenced by behaviorism and functionalism) and the traditional theological view of intelligence as a rational understanding and interpretation of reality. It stresses that human agency is fundamentally relational, rooted in the *imago Dei* (image of God), which involves intellect, free will, and stewardship. It discusses how AI, through practices like manipulative nudging, algorithmic governance, deskilling, and the *rapidification* of life, negatively impacts human freedom and promotes a paradigm focused on efficiency and control. Finally, the text proposes that the principles of Catholic social teaching, particularly subsidiarity and a focus on human dignity, offer ethical guidance for developing AI as a tool that genuinely supports human flourishing and responsible action rather than eroding it.

[1] Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, "*Antiqua et Nova*: Note on the Relationship Between Artificial Intelligence and Human Intelligence," January 28, 2025, §39, vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_en.html.

This publication invites us to revisit the traditional distinction between "freedom from" and "freedom for." This distinction is central to understanding both the threats AI poses to human agency and the constructive path forward. A freedom focused merely on overcoming external constraints can easily lead to a society where our choices are subtly manipulated or our skills are eroded by over-automation. In contrast, a freedom oriented toward the good calls us to develop the virtues necessary to act well, to seek genuine human connection, and to use our technological creations as tools for stewardship and service.

While AI applications can and do provide significant assistance—helping travelers navigate new places or empowering scientists to accelerate research—there is a growing sense that these same technologies are disempowering us in other, more subtle yet profound, ways. People feel trapped by automated systems, manipulated by algorithms, and deskilled by a growing dependence on apps. The promise of more time rings hollow in a world that feels increasingly accelerated, leaving us with longer to-do lists and a pervasive sense of loneliness. These anxieties reveal the inadequacy of a purely libertarian concept of freedom. What people truly seek is not just non-interference or an empty schedule, but a supportive community and a worthwhile purpose that makes action meaningful. This points to a more robust and constructive account of freedom: freedom for pursuing the good. Drawing upon the deep resources of Catholic thought, this book argues that proper human agency is not an undirected liberty but a "freedom for excellence." It is a freedom directed toward flourishing in right relationship with others and, most fundamentally, toward the love of God.

I cannot but repeat what I said in my preface to *Encountering AI:* This document "has been born from dialogue and is intended to facilitate further dialogue both within and beyond the Catholic world . . . it has been produced with a view to alerting a wider public to some of the fundamental questions on the meaning and purpose of human existence and the possible impacts of emerging technologies, questions that merit

greater attention and scrutiny." The commitment of Pope Leo XIV to put AI at the center of his mission, and the interest this commitment has evoked globally, makes this publication both timely and pertinent. It is my hope that *Reclaiming Human Agency in the Age of AI* will open up debate and empower a wider public to find a voice and contribute to the necessary civic and political consideration of the transformation being effected by technology.

I would like to finish by formally expressing my gratitude to the authors for their individual and collective efforts. It has been a privilege to see your work emerge. I am very conscious of the demands this project has made of you, and I would also like to acknowledge the support you have had from your academic institutions and your families.

Bishop Paul Tighe
October 3, 2025

CHAPTER 1

INTRODUCTION

"We have never empowered individuals this much," claims OpenAI's Sam Altman.[1] A fundamental selling point of AI (Artificial Intelligence) is the promise that it will help people do more or at least do more of the things that they want to do. Another commentator says, "with AI agents handling the busywork, we can focus more on connecting to our passions."[2] In the vision sold by tech entrepreneurs, AI's speed, efficiency, and skill at handling repetitive tasks and paperwork will free people to focus on more creative, productive, and fulfilling tasks, have more leisure, or simply complete more of their to-do lists. Instead of wasting time on a commute, people can begin their workday in a self-driving car. No longer do we have to format standard emails; AI will do it.

Boosters go even further than merely predicting the replacement of mundane tasks. Part of the empowerment they promise is that people will reach new levels of achievement by leveraging the abilities of AI applications.[3] Performance will improve as generative AI analyzes vast amounts of unstructured data found on the Internet and in business files

---

[1] Danny Fortson and Katie Prescott, "Open AI's Sam Altman: 'We're About to Empower People More than Ever Before,'" *The Times*, February 10, 2025, thetimes.com/business-money/technology/article/open-ais-sam-altman-were-about-to-empower-people-more-than-ever-before-h3qzfjmwj. See similar discussion in "Planning for AGI and Beyond," *OpenAI*, February 24, 2023, openai.com/index/planning-for-agi-and-beyond.

[2] WIRED Brand Lab, "The AI-Powered Future of Leisure Time," *Wired*, wired.com/sponsored/story/the-ai-powered-future-of-leisure-time.

[3] We define AI applications in this book broadly, including everything from older expert systems to newer generative AI and other tools that are commonly referred to as AI.

to generate new insights. Other commentators laud AI's ability to help artists create new songs, essays, and art by simply inputting a prompt. Today's imperative is: "Save time and achieve more with AI."[4] According to these narratives, AI expands both the pace and the breadth of human action, thus strengthening human agency.

Other entrepreneurs and technology analysts, however, do not merely claim that AI will strengthen human agency. They go further by considering AI itself as an agent. "Agentic AI" is one of the most hyped paradigms of tech development.[5] Rather than provide specific directions, users could give an "AI agent" broad goals that it could then achieve according to its own strategies. An AI agent would be granted control of its own software tools and Internet-connected apps to book travel reservations or schedule meetings. Such an AI program would seem to act in the world itself.

In these ways, tech companies promise that AI will improve the world and extend the human sphere of action, or at least the sphere of action of human artifacts. And there is some truth to this optimistic narrative. AI already is assisting many people to act. Translation and mapping apps can help travelers navigate new places with much greater ease, allowing them to find their ways through foreign streets and menus. AI applications have provided support for people with many kinds of disabilities, such as text-to-speech apps that allow people with blindness or dyslexia access to more written content. AI programs like Alphafold, whose creators won the 2024 Nobel Prize in Chemistry, empower scientists to accelerate research into deadly diseases. These are all significant contributions.

At the same time, these promises of increasingly broader spheres of action ring hollow for many people and in many situations. A widespread and growing feeling of disempowerment is also tied to AI. For example,

[4] "Save Time and Achieve More with AI," *Thomson Reuters*, August 22, 2024, thomsonreuters.com/en/insights/articles/save-time-and-achieve-more-with-ai.

[5] Mark Purdy, "What Is Agentic AI, and How Will It Change Work?" *Harvard Business Review*, December 12, 2024, hbr.org/2024/12/what-is-agentic-ai-and-how-will-it-change-work.

maps and translation apps assist travelers, but those same travelers may also find themselves stuck in endless loops with airline chatbots while knowing that, if they could only connect to a person, their reservation problems could be quickly solved.[6] AI can make navigating faceless bureaucracies even more challenging. There are also serious concerns about skill loss: People lose language or navigational abilities that they previously had, becoming dependent on the app.[7] They never develop some skills, like essay writing, in the first place.[8] Rather than opening up vast amounts of time, the quickening pace of life leaves people with ever lengthening to-do lists and feelings of a lack of time.[9] Even if we had time for more activities, we lack friends with whom to do them, as loneliness spreads across individual nations and the world at epidemic levels.[10] Or, people lack the resources with which to fulfill their basic needs, as the tech economy

[6] Hugh Gusterson, "Introduction: Robohumans," in *Life by Algorithms*, ed. Catherine Besteman and Hugh Gusterson (University of Chicago Press, 2019).

[7] Sam Schechner, "How I Realized AI Was Making Me Stupid—and What I Do Now," *Wall Street Journal*, April 3, 2025, wsj.com/tech/ai/how-i-realized-ai-was-making-me-stupidand-what-i-do-now-5862ac4d.

[8] E.g., Hua Hsu, "What Happens After A.I. Destroys College Writing?" *The New Yorker*, June 30, 2025, newyorker.com/magazine/2025/07/07/the-end-of-the-english-paper; James D. Walsh, "Everyone Is Cheating Their Way Through College," *New York*, May 7, 2025, nymag.com/intelligencer/article/openai-chatgpt-ai-cheating-education-college-students-school.html.

[9] Wei Jiang, Junyoung Park, Rachel Xiao, and Shen Zhang, "As AI's Power Grows, So Does Our Workday," *CEPR*, March 28, 2025, cepr.org/voxeu/columns/ais-power-grows-so-does-our-workday.

[10] Office of the US Surgeon General, "Our Epidemic of Loneliness and Isolation: The U.S. Surgeon General's Advisory on the Healing Effects of Social Connection and Community," *US Department of Health and Human Services*, 2023, hhs.gov/sites/default/files/surgeon-general-social-connection-advisory.pdf; World Health Organization Commission on Social Connection, "From Loneliness to Social Connection: Charting a Path to Healthier Societies," 2025, who.int/teams/social-determinants-of-health/demographic-change-and-healthy-ageing/social-isolation-and-loneliness.

accentuates inequality.[11] There is a sneaking suspicion that our actions are no longer fully our own, as exposés reveal how people are manipulated by social media algorithms.[12]

These are not entirely new observations. There are many excellent books and articles voicing concerns about topics like surveillance capitalism, algorithmic governance, and other AI-driven phenomena. We will draw on these works extensively in the coming chapters. While there are many criticisms of AI's effects on human agency, it can be difficult to find constructive accounts of human agency. For example, in defending private action, it is easy to slip from an analysis that walls individuals off from the influence of AI to one that walls us off from other people altogether. Frequently, this work promotes an ideal definition of freedom that tends to separate people from any kind of interference or responsibility. Such an account does not seem right, nor does it fully meet the concerns surrounding agency. What people are often seeking is not pure, isolated freedom, but a friend with whom to act. Not just noninterference, but a supportive institution. Not unbounded, undirected liberty, but a reason to do things: a good purpose that makes action worthwhile. Our rightful desires for freedom must be set within a broader vision.

Both to fully describe AI's threats toward human agency and to respond to said threats, we need a more robust, constructive account of human agency, how it operates, and its aims. The goal of this book is to offer such an account and to redescribe the problems of AI in light of it. This will allow us to develop better responses for these problems while still recognizing the real and growing positive contributions that AI can make to human agency. To do so, we turn to the Catholic tradition of thought

[11] There is a rapidly growing literature on AI and inequality, but for a good overview, see Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press, 2022).

[12] Adam Alter, *Irresistible* (Penguin, 2017); Shoshana Zuboff, *The Age of Surveillance Capitalism* (PublicAffairs, 2019); Paul Scherz, *Tomorrow's Troubles: Risk, Anxiety, and Prudence in an Age of Algorithmic Governance* (Georgetown University Press, 2022), chapter 8.

on the human person. This tradition gathers four millennia of thought and several strands of reflection (philosophical and theological, Greek and Jewish, scholastic and mystical) to develop a fuller understanding of human agency. Our argument is that this ancient wisdom has the resources to confront and help solve novel contemporary problems.

Intelligence and Agency, Human and Artificial

Given the complicated connection between agency and intelligence, it will be helpful to start by considering the concept of intelligence itself. In our previous book, *Encountering Artificial Intelligence*, we examined this issue extensively,[13] so here we will be brief. To claim that a program is "Artificial Intelligence," researchers need a definition of intelligence. We noted that these definitions often share characteristics: "First, rationality and understanding become the *logical manipulation* of symbolically represented information; and second, intelligence becomes *efficacious problem-solving*."[14] We see this, for example, in Russell and Norvig's classic textbook *Artificial Intelligence: A Modern Approach*, where they distinguish historical approaches to AI as focusing on thinking or acting.[15] Because thinking is internal and action is external, action is the more measurable focus, the one that can actually be worked on. For this reason, Russell and Norvig state that their approach will be based on "rational action": "A rational agent is one that acts so as to achieve the best outcome."[16] We will discuss this further below.

It is worth noting that this approach is in contrast to the traditional Western philosophical and theological approach to intelligence as the

[13] AI Research Group for the Centre for Digital Culture of the Dicastery for Culture and Education of the Holy See, *Encountering Artificial Intelligence: Ethical and Anthropological Investigations*, ed. Matthew J. Gaudet, Noreen Herzfeld, Paul Scherz, and Jordan J. Wales (Pickwick Publications, 2024), 15–17, 21–23, 57–68.

[14] AI Research Group, *Encountering Artificial Intelligence*, 59.

[15] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall, 2003), 1–5.

[16] Russell and Norvig, *Artificial Intelligence*, 4.

experience of rational understanding, which is not primarily about calculation or problem-solving or agency, but primarily about interpreting reality as it truly is.[17] It is about the human mind grasping at the world around us and actually "*getting it*," "catching the drift," "seeing the point." The next step of agency moves toward external action. But experiencing understanding comes first, and then agency follows as a response to understanding.

Russell and Norvig recognize this difference and acknowledge that other AI researchers have tried to take more human-centric or thinking-centric approaches, but that these are not easily scientifically investigable nor technologically constructible. Therefore, the focus of AI must be on rational action, since it is the most tractable approach.[18] This tactic has behaviorist influences insofar as it focuses on outward appearance and not inward experience. It has functionalist influences because it concerns achieving goals. And this is an approach which is above all practical: Actions must be taken and objectives met. Theoretical questions such as

---

[17] AI Research Group, *Encountering Artificial Intelligence*, 58, citing Boethius, "A Treatise Against Eutyches and Nestorius," in *The Theological Tractates*, trans. Hugh Fraser Stewart (Heinemann, 1918), 85; Jordan Joseph Wales, "Participatory Spiritual Intelligence: A Theological Perspective," in *Perspectives on Spiritual Intelligence*, ed. Fraser Watts and Marius Dorobantu (Routledge, 2024); Brian Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (MIT Press, 2019); Josef Pieper, *Leisure: The Basis of Culture* (Pantheon, 1952), 33–34.

[18] Russell and Norvig, *Artificial Intelligence*, 2–5. This is a humbler approach than that of Alan Turing, who famously stated, "We may now consider the ground to have been cleared and we are ready to proceed to the debate on our question, 'Can machines think?' . . . I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent. chance of making the right identification after five minutes of questioning. The original question, 'Can machines think?' I believe to be too meaningless to deserve discussion. Nevertheless 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. I believe further that no useful purpose is served by concealing these beliefs." See Alan M. Turing, "Computing Machinery and Intelligence," *Mind*, New Series 59, no. 236 (1950): 442.

the "experience of reality" are not within the realm of investigation—hence the mystery that concerns some researchers about whether AI "feels" anything or is conscious, and the ethical worries that accompany that concern.[19]

Much of contemporary artificial intelligence research, then, is framed as rational action because that is what researchers have the tools to investigate. But from an outsider's perspective that sees intelligence as the experience of understanding reality more than just action that reasons toward an objective, this AI framing falls short. This is not to blame AI researchers; they are stipulating a technical definition for particular tasks. It is merely to point out that confusing these definitions can be extremely significant.

Given this background on intelligence, the next move is to look at agency. We can start by contrasting human agency with the full Russell and Norvig definition of agency in computer science:

> An **agent** is just something that acts. . . . But computer agents . . . have other attributes . . . such as operating under autonomous control, perceiving their environment, persisting over a prolonged time period, adapting to change, and being capable of taking on another's goals. A **rational agent** is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome.[20]

---

[19] See, for examples, Patrick Butlin, Robert Long, Eric Elmoznino, et al., "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness," *arXiv.org* (August 22, 2023), arxiv.org/abs/2308.08708; David Gamez, "The Potential for Consciousness of Artificial Systems," *International Journal of Machine Consciousness* 1, no. 2 (2009): 213–223, worldscientific.com/doi/abs/10.1142/S1793843009000190; Angela Langdon, Matthew Botvinick, Hiroyuki Nakahara, Keiji Tanaka, Masayuki Matsumoto, and Ryota Kanai, "Meta-learning, Social Cognition and Consciousness in Brains and Machines," *Neural Networks* 145 (2022): 80–89, doi.org/10.1016/j.neunet.2021.10.004; Wanja Wiese and Karl J. Friston, "AI Ethics in Computational Psychiatry: From the Neuroscience of Consciousness to the Ethics of Consciousness," *Behavioural Brain Research* 420, no. 113704 (2022) doi.org/10.1016/j.bbr.2021.113704—all of which are cited in Louie Kangeter and Brian Patrick Green, "AGI and Slavery," *AI Ethics* 5 (2025): 3312, link.springer.com/article/10.1007/s43681-024-00618-z.

[20] Russell and Norvig, *Artificial Intelligence*, 4.

This definition focuses entirely on doing. It does not matter how the agent comes to the decision for action, only that it does so autonomously. The aim of agency is to achieve an optimal outcome, but the language of "best expected outcome" indicates a very particular understanding of the good. That framing indicates that the good is understood in terms of utilitarian rational actor theory. As we will discuss in chapter 4, the idea of rational actor theory is that the best action can be mathematically calculated. While this definition is fine for the particular needs of computer science, it is a very restricted idea when it comes to a definition of human agency.

This tie between agency and action is acceptable, although, drawing from the Aristotelian tradition, the notion of action should extend beyond mere doing. Aristotelians understand even simply being alive as an "act," so all people have agency, even those who lack the exercise of consciousness, rationality, or the ability to act physically in the world. Our focus in this book is on a more particular form of agency: *responsible* human agency. It is this responsibility that Pope Francis argued was threatened by AI:

> We would condemn humanity to a future without hope if we took away people's ability to make decisions about themselves and their lives, by dooming them to depend on the choices of machines. We need to ensure and safeguard a space for proper human control over the choices made by artificial intelligence programs: human dignity itself depends on it.[21]

By "responsible agency," we mean voluntary action that is freely chosen, informed by reason, and aimed at achieving human goods. Ultimately, it aims at human flourishing, which is achieved through right relationship with God and others. In contrast to the computer-science definition of

[21] Francis, "Address of His Holiness Pope Francis to the G7 Session on Artificial Intelligence," Borgo Egnazia (Puglia), June 14, 2024, vatican.va/content/francesco/en/speeches/2024/june/documents/20240614-g7-intelligenza-artificiale.html.

agency, responsible human agency relies on deliberation about reasons and conceptions of the good. It depends on the ability of the person to responsibly reflect on goods that they rationally grasp and then choose to pursue them. The person comes to a judgment about action that they then enact. The Vatican document *Antiqua et Nova* argues that

> Between a machine and a human, only the human can be sufficiently self-aware to the point of listening and following the voice of conscience, discerning with prudence, and seeking the good that is possible in every situation.[22]

As chapters 2 and 4 will explore in depth, AI, at least in any foreseeable form, lacks the ability to interpret and judge the good, and cannot freely seek it.

Reflection on the good is important because proper human agency cannot be understood in terms of a merely libertarian freedom *from* external constraint or influence. Instead, it is a freedom *for* pursuing the good, a freedom *for* excellence.[23] Human agency is given to us so that we might pursue the purpose of human life, which is to live in flourishing relationships with others. In the Christian tradition, it is most fundamentally a freedom that enables the love of God. Yet, to flourish means developing the character traits necessary to act well. We also have to resist temptations to misinterpret our good to be something like, for example, generating economic value. Thus, it is a freedom directed toward

[22] Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, "*Antiqua et Nova*: *Note* on the Relationship Between Artificial Intelligence and Human Intelligence," January 28, 2025, §39, vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_en.html.

[23] Servais Pinckaers, *The Sources of Christian Ethics,* 3rd ed., trans. Mary Thomas Noble (Catholic University of America Press, 1995), 354–378.

excellence and virtue, which must be supported by and shared with others to achieve its fullest form.[24]

Human agency is thus an essential characteristic of the kind of being that we are. The radical potential for responsible agency is a part of human nature and helps us to fulfill our purpose.[25] Because of this tie to the very nature and purpose of the human person, undermining or replacing human agency is an attack on human dignity.[26] Agency is merely one aspect of human dignity, but it is the one that we address here, as it seems to be especially endangered by AI. Moreover, when people lose their sense of agency, they can lose their confidence that they can build a better shared future. Thus, defending responsible human agency is essential both to respecting our intrinsic dignity and to pursuing our future flourishing. This book will also concentrate on AI systems as externally imposed risks to human agency, not AI use as a voluntarily chosen reduction of human agency. This distinction is important: When we turn driving over to an autonomous vehicle, we are effectively using our agency to choose to reduce our agency. We choose to put ourselves at the mercy of a machine, for the sake of some other good, the judgment and justification of which is another question.[27] Given the present context, external risks to agency

[24] According to Christian teaching, we require grace from God and the transformation enabled by Jesus Christ's sacrificial death for us to fully seek the good. God became human so that we might be free of our deformed, selfish inclinations in order to turn to others: "For you were called for freedom, brothers. But do not use this freedom as an opportunity for the flesh; rather, serve one another through love" (Galatians 5:13).

[25] As we will explore more fully in later chapters, it is the *potential* for agency that is part of human nature, but plenty of people do not actualize this potential at any one time for any number of reasons: they are asleep, in a coma, too young, have a disability, suffer from an illness, etc. This failure to actualize agency impairs neither their dignity nor their fundamental potentiality for agency.

[26] Dicastery for the Doctrine of the Faith, *Dignitas Infinita*, 2024, §1–32, vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20240402_dignitas-infinita_en.html.

[27] Fully autonomous weapons are perhaps the ultimate means to attempt to reject responsibility and prevent guilty consciences as well. The Church condemns autonomous weapons, for example, in Dicastery for the Doctrine of the Faith and Dicastery for Culture

are more than enough for one book and voluntarily choosing to reduce agency must be a topic for the future.

Plan of the Book

We will begin by describing the Catholic account of human agency, where human agency is both individual and social. Chapter 2 will examine the individual faculties and purposes that make up human agency. Chapter 3 will explore the effects of other people, tradition, and society on individual agency, asking about both the good and the bad effects of our social dependence on others. These chapters are fundamentally grounded in the scripture and tradition of the Catholic Church, in a framework shared with many other Christian denominations. Additionally, this account developed in conversation with Greco-Roman philosophy and other religious traditions, so we believe it is also recognizable to those coming from outside the Christian tradition.[28]

In chapter 4, the book will turn specifically to AI, arguing that, at least as it currently exists and under the current major paradigms of development, AI does not possess responsible human agency. Primarily this is because it lacks the natural capacity to consciously reflect on the meaning and direction of its actions. We can better understand why people mistake some AI programs for human-like agents once we see that commentators frequently rely on mistaken understandings of human agency.

Parts II and III are the applied sections of the book. Part II looks at how AI can deform human agency in different areas, such as governance, commerce, and personal life. It examines the specific techniques and

---

and Education, *Antiqua et Nova*, §100, quoting Pope Francis: "No machine should ever choose to take the life of a human being."

[28] For a broader discussion of this understanding of the universality of an ethics based in a shared understanding of the human person, see International Theological Commission, "In Search of a Universal Ethic: A New Look at the Natural Law," 2009, vatican.va/roman_curia/congregations/cfaith/cti_documents/rc_con_cfaith_doc_20090520_legge-naturale_en.html.

aspects of AI that threaten agency: nudging and surveillance capitalism in chapter 5, deskilling and algorithmic governance in chapter 6, and misinformation and rapidification in chapter 7. This analysis of the threats of AI to human agency will lead us to develop a set of responses in Part III based in Catholic social teaching (CST), as described in chapter 9. These responses include regulatory constraints on AI that can protect human agency, discussed in chapter 10. Chapter 11 also suggests more constructive frameworks for programming, organizing, and implementing AI that would encourage forms of AI that foster human agency. These responses will allow people to use AI in ways that support rather than undermine agency.

We believe that ultimately AI can be a real aid to human agency and flourishing. Today, however, it is contributing to a broader malaise regarding human possibilities. Too often, it seems to disempower. Yet that is not a necessary outcome of this technology. Through better design and more just social structures, our prudent use of AI applications may indeed empower people, but in a different and more profound way than that imagined by Sam Altman. Rather than simply freeing up our time or increasing our efficiency, AI can promote human agency in more genuinely fulfilling ways; it can fulfill its potential to protect and advance human dignity.

Part I

Understanding Human Agency

An accurate description of the problems that AI raises for human agency, and appropriate solutions to those problems, requires a clear understanding of what human agency is. The next three chapters discuss aspects of human agency and how it differs from the seemingly agentic powers of AI applications. Chapter 2 grounds agency in the Christian understanding of human creation in the image of God. Part of this image consists in human nature's radical capacities for reason and free will, which are necessary for the rational deliberation of responsible agency. Though grounded in the capacities of each individual, due to our relational nature, human agency develops in and depends upon other individuals and a broader social environment. These influences can shape our agency both for good and for ill, as chapter 3 discusses. Chapter 4 argues that the interpretive and deliberative aspects of human agency distinguish it from the capacities that AI possesses. Agency is ascribed to AI applications only due to misunderstandings and impoverished theories of human agency. The account offered in these chapters paves the way for a detailed discussion in Part II of the problems AI raises for human agency.

Chapter 2

Responsible Agency in Catholic Thought

Pope Francis proposed that "to speak of technology is to speak of what it means to be human and thus of our singular status as beings who possess both freedom and responsibility."[1] Thus, we cannot fully grasp the scope and importance of the responsible agency at the heart of many current debates over the technology of AI unless we have a better understanding of the nature and purpose of human beings. In this chapter, we will explore a Christian view of the human person, what is called philosophical or theological anthropology.

The beginning of the book of Genesis provides the foundation for the Catholic understanding of the human person: "God created mankind in his image / in the image of God he created them / male and female he created them" (Genesis 1:27). Each subsequent person is also created "in the image of God," also known as the *imago Dei*. Despite the centrality of the *imago Dei* to Christian theological anthropology, its meaning has been heavily debated among theologians across the ages. In part, this is because it, like many important concepts from scripture, has rich layers of meaning. First, the concept of the *imago Dei* is a marker of humanity's special place in creation, as steward of all that was created, intended by God for a unique relationship that participates in God's own life, and as participant in the ongoing creation. The *imago Dei* also relates to human dignity: "To be created in the image of God means to possess a sacred value

[1] Francis, "Address of His Holiness Pope Francis to the G7 Session on Artificial Intelligence."

that transcends every distinction of a sexual, social, political, cultural, and religious nature."[2] Yet, it also reflects human qualities, such as the "gift of intelligence."[3] As Pope St. John Paul II described, "The biblical author sees as part of this image not only man's dominion over the world but also those spiritual faculties which are distinctively human, such as reason, discernment between good and evil, and free will."[4] This capacity for reason is to be used for "responsible care," understood as governing "the world with wisdom and justice."[5] Technology is an important aspect of the stewardship by which we enact God's directive to "till and keep" the earth.[6] This concept of the person also reveals humans as fundamentally relational: Created in the plural, we are ordered toward communion with others, and learn from and collaborate with others to act for the good. Finally, the *imago Dei* points us to our last end of communion with God and formation in the virtues that direct us to that end. Because of the *imago Dei*, we have the capacity for free action in accord with the good that we choose to love.[7] Responsible agency is essential to realizing our true, ultimate good. As the Second Vatican Council taught, "Authentic freedom is an exceptional sign of the divine image within man . . . so that he can seek his Creator spontaneously, and come freely to utter and blissful perfection through loyalty to Him."[8] As a precursor to our exploration of AI and its relationship to agency, this chapter more fully explores the image

---

[2] Dicastery for the Doctrine of the Faith, *Dignitas Infinita*.

[3] Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §1.

[4] John Paul II, *Evangelium Vitae*, 1995, §34, vatican.va/content/john-paul-ii/en/encyclicals/documents/hf_jp-ii_enc_25031995_evangelium-vitae.html.

[5] Second Vatican Council, *Gaudium et Spes*, 1965, §34, vatican.va/archive/hist_councils/ii_vatican_council/documents/vat-ii_const_19651207_gaudium-et-spes_en.html.

[6] Genesis 2:16. See Francis, *Laudato Si'*, 2015, §66, w2.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html.

[7] International Theological Commission, "Communion and Stewardship: Human Persons Created in the Image of God," 2004, vatican.va/roman_curia/congregations/cfaith/cti_documents/rc_con_cfaith_doc_20040723_communion-stewardship_en.html.

[8] Second Vatican Council, *Gaudium et Spes*, §17.

of God, intellect and free will, stewardship through technology, and relationality.

In explicating the *imago Dei*, we will draw particularly on the thought of St. Thomas Aquinas (d. 1274), who made a singular contribution to synthesizing preceding strands of thought on theological anthropology. In drawing together a variety of classical, Jewish, Christian, and Muslim philosophical sources, his work became a major resource for later magisterial and theological reflection. With only a brief space available, we will zero in on those aspects of the *imago Dei* that relate to responsible agency. Thus, it will be a partial exploration of the idea of the image of God, and important themes, especially Jesus Christ as the true image, will not be explored as fully as they could.

Imago Dei

Central to the Christian vision of the person is the understanding that human beings do not make themselves; we are created. This dependence does not diminish our dignity but allows us to bear the imprint of our Creator. As Henri de Lubac argues, this message was liberating for early Christians at a time when many pagans believed themselves to be subject to and threatened by ruthless and dangerous powers and fates. Christians, along with some Hellenistic philosophical schools like the Stoics and Platonists, instead understood their world to be governed by a rational, caring providence. They knew themselves elevated above determination by worldly powers and the contingencies of history because of our creation in God's image and likeness.[9]

Beyond the dignity implied by bearing God's image, the *imago Dei* also implies something about our capacities. God the Creator is pure spirit (John 4:24). Thus, the *spiritual* faculties of the human person, such as intellect and free will, are essential ways that humans are in the image and likeness of God. As God's providence governs creation, humans preside

[9] Henri de Lubac, *The Drama of Atheist Humanism* (Ignatius Press, 1995), 19–25.

over the embodied, material aspects of our own being, which are also essential to us as "rational animals," since the person is an inseparable unity of body and soul.[10] Through our bodies and the work of our hands, humans experience the gift of creation, are cared for by God's providence, and extend the governance of embodiment to the rest of creation in stewardship (Genesis 1:26-30).

In a specifically Christian context, the concept of the image and likeness of God also invokes the Trinitarian character of divine personhood. As Aquinas explains it, the divine Persons are one in nature but distinct by mutual relations. This is suggestive for the nature of human persons as well. Though our shared nature does not make us one in being or "consubstantial," as are the divine persons, we are meant to be in relationships of communion with others. Relations come in different forms (spatial, temporal, causal, logical, and so on). The relations characteristic of persons, however, are relationships that depend on exercising those features particular to the image and likeness of God: intellect and will. Those faculties that we share with God allow us to transcend ourselves and see ourselves with the eyes of another.[11] These spiritual faculties render us open to relationships that are characteristic of persons, relationships of mutual knowledge and love which result in communion. As Aquinas says, following Augustine, the kinds of relations that constitute the divine persons flow from the intellect and will, respectively: The Son as the *Logos* is conceived in the mode of the intellect, the Spirit as Love in the mode of the will.[12]

This relationality and our spiritual faculties are grounded in human nature. Just as the divine relationality is rooted in divine nature, so it is for

---

[10] See the recent discussion in Dicastery for the Doctrine of the Faith, *Dignitas Infinita*, §18.

[11] This sentence draws on Helmuth Plessner, *Levels of Organic Life and the Human: An Introduction to Philosophical Anthropology*, trans. Millay Hyatt (Fordham University Press, 2019).

[12] This means that there is no contrast between an outlook that roots the image and likeness of God in the intellect and will as spiritual faculties of the soul and one that explains it in terms of relationality; false dichotomies must be avoided.

us with human nature and relationality. As a consequence, the absence of an unfolded, or fully actualized, rationality does not mean the absence of personhood.[13] Even if persons cannot actualize their rational nature because of disability, or because they are still in their incipient stages in the womb, or because they are in a "vegetative state," they are still persons, sharing in the image of God. It is their nature as human persons with a rational soul that gives them dignity, not the extent to which they are able to exercise these faculties.[14] Likewise, it is important to recall this at a time

[13] Such claims have been made, e.g., by Eberhard Jüngel, "Hoffen, Handeln - und Leiden. Zum christlichen Verständnis des Menschen aus theologischer Sicht," *Evangelische Kirche in Deutschland,* January 28, 2002, ekd.de/bioethik_juengel_vortrag_020128.htm. Invoking the personal relationship of recognition, in particular, can have this result, as can be seen in Alberto Giubilini and Francesca Minerva, "After-Birth Abortion: Why Should the Baby Live?" *Journal of Medical Ethics* 39 (2013): 261–263. But persons are not constituted by the relation of recognition; rather, they *demand* recognition based on what they are antecedently.

[14] For a deeper exploration of these ideas, see, for example, the work of Miguel Romero, who argues that even if a person does not exhibit the same capabilities of enacting reason as others, this does not mean that they do not possess an intellectual nature. This is not only because the *imago Dei* is intrinsic to a person but also because the person's intellectual nature is not limited to the actualization of rationality alone. "Strictly speaking, for Aquinas, human nature is not 'rational,' but 'intellectual'—the deliberative use of reason, of course, being only one of the acts of the intellectual soul that distinguishes the human being from other animals. . . . As God knows and loves God's self, the human being reflects the image of the Creator when it knows and loves that which we share in common with the Creator: our intellectual nature." See "The Happiness of 'Those Who Lack the Use of Reason,'" *The Thomist* 80, no. 1 (January 2016): 57. For a comparison and contrast between an understanding of disability according to Thomas Aquinas and Edith Stein, see Mariele Courtois, "Biomedical Challenges to Identity and Parenthood: An Investigation into the Ethics of Genetic Technologies at the Beginning of Life" (PhD diss., The Catholic University of America, 2022). Acknowledging this distinction helps us to understand how all persons—no matter their individual rational capacities—can freely take part in the moral life with God. It is for this reason that John Paul II can clearly state that: "The world of rights cannot only be the prerogative of the healthy. . . . The disabled are not different from other people, which is why, in recognizing and promoting their dignity and rights, we recognize and promote our own dignity and rights and those of each one of us." See "Message of John Paul II on the Occasion of the International Symposium on the Dignity and Rights of the Mentally Disabled," January 5, 2004, §5, vatican.va/content/john-paul-ii/en/speeches/2004/january/documents/hf_jp-ii_spe_20040108_handicap-mentale.html.

when intelligence is conceived in terms of performance, something in which AI may come to easily outpace us, without therefore gaining the status of personal dignity.

Distinctly Human Capacities of Rational Nature: The Intellect and the Free Will

What we have said so far still leaves the question of what exactly reason and free will are and why we were given them. Their importance emerges from their purpose and their relationship to the *imago Dei*. The human person is created in the *imago Dei* simply because of the special love that God has for humans, derived from God's divine nature and God's intention to be in a uniquely close relationship to us. To participate in this relationship, the human being identifies and values the gifts of God as well as freely chooses to respond to and love them. God desires a response of love that is genuine, free, and aware. Thus, humans are gifted particular capacities—an intellect and a will—that are unique to persons, who can freely direct their actions and reveal their loves and values in relationship with others. The human rational nature provides the human person with the ability to discern the truth through the intellect and the ability to choose the good through free will.

The Intellect

The intellect makes it possible for the human person to understand and assent to the truths they encounter. The human intellect grasps both speculative (theoretical) and practical truths. A distinction between these activities does not imply that they are disconnected or irrelevant to each other. In fact, each relies upon the other; to act well in the world, we must understand the surrounding world and the meaning and value of our actions.[15] Likewise, the theoretical intellect only develops through

[15] Thomas Aquinas writes, "But perfection and rectitude of reason in speculative matters depend on the principles from which reason argues; just as we have said above that science depends on and presupposes understanding, which is the habit of principles. Now in human

deliberately undertaken action. The tie between intellect and agency will be important for later discussions of the distinction between AI and humans, as well as the effects of AI on our information ecologies.

Within the practical intellect, Aquinas also draws a distinction between moral activity and what he calls artistic activity. Moral activity pursues the good of the human being. "Art" has a broader meaning for Aquinas and his contemporaries than for us, suggesting any activity whose goal is to produce something external to the agent. It is more like our idea of craft and is a translation of the Greek term *technē*, from which the English word "technology" is derived. While human productive activity is important and can be used in such a way to reflect and delight in order and beauty as well as to teach and celebrate truths, art itself does not ensure that the actor is seeking the ultimate happiness or the true goods of a flourishing life. One can use medical skills to poison as well as to heal, or technical skills to destroy as well as to create. For this reason, art "is called a virtue. And yet it falls short of a perfect virtue, because it does not make its possessor to use it well; for which purpose something further is requisite: although there cannot be a good use without the art."[16]

Thus, the activity of making, though a significant human feat, does not involve the most valuable human act. There is greater perfection in human acts that embody a full understanding and assent toward the perfect end to which humans are ordered.[17] Intellectual apprehension of the truth and values of the world is a prerequisite for a human to freely make meaningful activity and a gift of self within the world. The practical and speculative

---

acts the end is what the principles are in speculative matters, as stated in *Ethic.* vii, 8. Consequently, it is requisite for prudence, which is right reason about things to be done, that man be well disposed with regard to the ends: and this depends on the rectitude of his appetite." See *Summa Theologica* [*ST*] I-II, q. 57, a. 4, trans. Dominicans of the English Province (Benziger Bros., 1947).

[16] Aquinas, *ST* I-II, q. 57, a. 3, ad 1.

[17] "The consequence is that more praise is given to a craftsman who is at fault willingly, than to one who is unwillingly; whereas it is more contrary to prudence to sin willingly than unwillingly, since rectitude of the will is essential to prudence, but not to art," *ST* I-II, q. 57, a. 4.

intellects are foundational to the human ability to create, to be in relationship, and to serve our vocations. This interconnection allows the worker's creativity and products to bear the "impress of his personality"[18] and to "speak of their authors; they enable us to know their inner life."[19] Creation reflects the love, conviction, and understanding of its Creator. Thus, developing technologies like AI can be valuable expressions of the person, but only if they are directed toward a greater good.[20]

Free Will

Rational faculties and their acts are also presupposed in acts of *free will*. We cannot love what we do not know; we cannot choose a thing of which we are ignorant.[21] Through rational perception, the objects of our choice become present in us. They are moved into our minds before we can move outward again for the sake of action.[22] As Aquinas explains it, following most of ancient philosophy, all people are motivated to choose the good and, ultimately, their happiness in a flourishing life. The happiness of a rational creature consists in God, who is unlimited, subsisting goodness, universal in containing all goodness in himself. However, in everyday choices, people are rarely presented with a clear choice for or against flourishing and the good. Instead, we are presented with choices among many particular goods, each of which may or may not direct us toward the ultimate good. Should I take this job because of its higher pay or another job that will allow greater creativity? Should I give this money to this

[18] Leo XIII, *Rerum Novarum*, May 15, 1891, §9, vatican.va/content/leo-xiii/en/encyclicals/documents/hf_l-xiii_enc_15051891_rerum-novarum.html.

[19] John Paul II, "Letter to Artists," 1999, §2, vatican.va/content/john-paul-ii/en/letters/1999/documents/hf_jp-ii_let_23041999_artists.html.

[20] See Brian Patrick Green, "Ethics Is More Important than Technology," Markkula Center, August 10, 2020, scu.edu/ethics/all-about-ethics/ethics-is-more-important-than-technology/.

[21] Saint Augustine, *De Trinitate*, 10.1., ed. John E. Rotelle, trans. Edmund Hill (New City Press, 2012).

[22] Aquinas, *ST* II-II, q. 27, a. 4. The will is *voluntas in ratione*; Aristotle, *De an.* III, 9 (432 b 5–6). Both emanate directly from the same spiritual substance, but the will presupposes the intellect. Aquinas, *De veritate*, q. 22, a. 10, ad 2 and a. 11 ad 6; also, *De veritate*, q. 23, a. 1c.

homeless person or to a charity or use it to buy a present for my spouse? People have to ask whether any particular good leads us to a more encompassing goodness or not. In other words, people deliberate freely among a range of possible choices. Our rational faculties give us the ultimate goodness as our *goal*, but they are also applied in discerning the *means* toward this end.[23]

Such deliberation is presupposed in free choice; it is what "specifies" our free choice.[24] Free acts are not random; they are not the mere indeterminacy of chance events or the result of random generators, as in some machine simulations of agency. Blind chance is the opposite of informed consent. But if we know what we are doing, then we can responsibly own our choice and be free in it.

This is not yet the whole story of free choice, however. After the deliberations of reason have run their course, we still have to make a choice. Merely *knowing* the good is not a choice. We still have to *perform* or "exercise" the choice through our will, and we can fail to do so.[25] Unfortunately, almost all of us have failed to do what we know the right action is, whether from fear or greed or laziness.

Here is where genuine spontaneity, freedom, and agency are exercised. As it presupposes the acts of rational knowledge, such agency requires the understanding provided by our intellect. Chapter 4 will discuss how this differentiates human agency from any kind of agentic AI. Human agency is a spontaneity or self-movement that shares in the creative activity of God, who begins things in ever-new freshness. Our own acts of making

[23] Aquinas, *ST* I, q. 82, a. 1 ad 3, a. 2; q. 83. Also, *ST* I-II, q. 13, a. 3; *De veritate* q. 24, aa. 1–6. Tilman Anselm Ramelow, "Nachwort," in *Thomas von Aquin: Über die Wahrheit - De veritate*, Teilband 5 (Q. 21–24) (Meiner, 2013), 311–399, esp. 358–363.

[24] Aquinas, *De Malo*, q. 6; *ST* I-II, q. 9, a. 1.

[25] Aquinas, *De Malo*, q. 6; *ST* I-II, q. 9, a. 1.

share in this newness.[26] AI can only mimic this natality by remixing preexisting patterns learned by Large Language Models, or LLMs.[27]

Making

Humans use our intellect and free will in the activities of creating, producing, and working. Animals without a rational nature are certainly able to demonstrate the capacity to build. We see this in the achievements of birds in building nests, beavers in building dams, and bees in building hives. These are complex structures that also reflect the order and generosity of the divine mind who is the source of being of each of these creatures. Nevertheless, rational creatures can create not simply out of a response according to instinct, but rather from a deliberate act of the will. Human creativity is properly ordered when we freely act according to a conviction of truth and bestowal of love. The creative activity of human beings is an act of the free will and is one unique way that human beings can reflect the activity of their own Creator. These creative acts are a gift by which God allows human beings to have a greater participation in God's own knowledge and care from within grace-filled acts. Creative acts are thus rational acts: They exhibit human imagination and ingenuity and correspond to an understanding of the ordered world. Creative products should serve the truth, beauty, and goodness given to us in the world. Human creative acts are called to a higher meaning. They are meant to follow from the individual's vocational calling to assist with the redemptive plan of God. They are not random outcomes of an irrational stumbling or simply the chance results of the march of progress but are to be intentional, loving, truthful, and meaning-filled acts drawing forth a

[26] For a secular development of this idea of the human ability to create something new, see the discussion of natality in Hannah Arendt, *The Life of the Mind* (Harcourt Brace Jovanovich, 1978), 29, 109; Hannah Arendt, *The Human Condition* (University of Chicago Press, 1998), 177; Anne O'Byrne, *Natality and Finitude* (Indiana University Press, 2010), 78–106; Scherz, *Tomorrow's Troubles*, 50.

[27] A. Ramelow, "Can Computers Create?" *Evangelization and Culture* 1 (2019): 39–46; Scherz, *Tomorrow's Troubles*, 50–53.

change of heart, deeper love, and fulfilling inspiration from those who encounter these creative works—whether they be artistic masterpieces or technological feats.

An important distinction must be drawn, however, between human and divine creative acts. In his "Letter to Artists," John Paul II emphasized the distinction between a creator and a craftsman:

> The one who creates bestows being itself, he brings something out of nothing—*ex nihilo sui et subiecti*,[28] as the Latin puts it—and this, in the strict sense, is a mode of operation which belongs to the Almighty alone. The craftsman, by contrast, uses something that already exists, to which he gives form and meaning. This is the mode of operation peculiar to man as made in the image of God. . . . God therefore called man into existence, committing to him the craftsman's task. Through his "artistic creativity" man appears more than ever "in the image of God", and he accomplishes this task above all in shaping the wondrous "material" of his own humanity and then exercising creative dominion over the universe which surrounds him.[29]

To whatever extent we can acknowledge and celebrate the human ability to create, we must do so with immense humility toward the restricted, limited nature of our creativity in comparison to the divine ability to cause existence itself. Not only does God bring forth being, but God also gives purpose to it—a purpose that relates back to God. God as ultimate source and purpose of all of creation is a truth by which to inform our interactions with the products of our making. People should intentionally strive to use the products of making to deepen human relationships with God.[30]

---

[28] We translate this as: "Out of nothing of Himself or of another substance."

[29] John Paul II, "Letter to Artists," §1.

[30] On the integration of human and divine making and its problematic history in practice and theory, see Anselm Ramelow, "Technology and Our Relationship with God," *Nova et Vetera* 22 (2024): 159–186.

Agency as Relational

Though the *imago Dei* is tied to the human capacity for reason and free will, as the last section argued, this agency cannot be understood in individualistic terms alone since human agency is, at its heart, relational. Every individual's agency to freely seek happiness, to deliberate upon and choose actions, must be respected. Ultimately, agency is ordered to seeking God. However, we reach God through our neighbors, through communal action and acts of service. We serve God by serving the least among us (Matthew 25:45). It is for this reason that the Good Samaritan enacts the paradigmatic form of Christian agency by taking upon himself the duty to care for a person in need (Luke 10:25–37). By serving a person who has no obvious call on him other than shared humanity, the Samaritan mirrors God's gracious salvation of humanity through the Incarnation of Jesus Christ.

Because of our relational nature, God chooses to save humanity through the medium of community. God calls the people of Israel in the Hebrew scriptures. Jesus Christ gathers a group of disciples about himself. The Church comes forth from the sending of the Holy Spirit. Further, the community of the Church is bound together in the body of Christ, so that the good of any redounds to the good of all (1 Corinthians 12:26).

We need not only refer to scripture to understand that human agency only attains its end of a flourishing life through communal action. Indeed, as Aristotle and other ancient authors realized, humans are political animals.[31] The metaphor of a community as a body was also a staple of ancient Roman political thought.[32] Though a person may be able to obtain the bare means of life through their own efforts, it is only through community that people attain the good life, including richer forms of fulfillment like education and crafts. Yet we seek community not only because of its pragmatic importance but also because of the good of being

[31] Aristotle, *Politics*, trans. Carnes Lord (University of Chicago Press, 2013), I.2.

[32] Livy, *History of Rome*, vol. 1, trans. B. O. Foster (Harvard University Press, 1919), 2.32.8–12.

in relationship. Relations of friendship and solidarity are themselves essential for human flourishing.

These dual aspects—people both need community to achieve their ultimate ends and delight in the good of being in relationship itself—point to the idea of the common good.[33] Every association, be it a family, a school club, a labor union, a small business, a local charity, or even a national government, has both an extrinsic end that can only be achieved through the joint work of many people and an intrinsic end that comes from the good of those people being in association. This intrinsic good cannot be appropriated by any single member because it arises out of the good of joint action. The school's extrinsic end is educating students, but the school community also allows for the good of a shared life of the mind. The business has the extrinsic end of producing some product, but it also has a good of shared work toward an end. A common good is therefore a good that cannot be attained by an individual alone. It requires the action of many people. Because part of the human good arises from this kind of joint agency of working together, social associations ought to be promoted in a good society.

Of course, cooperative action does not scale uniformly. Large-scale cooperative action, by its nature, often loses the direct interpersonal knowledge and relationship that is the intrinsic good of a smaller operation. The CEO does not personally know the average worker in an international corporation, and nationally elected representatives or heads of state do not personally know their constituents. Rather, interpersonal relationship is often replaced by impersonal rules, orders, hierarchies, and other structures as organizations scale up. In order to maintain some level of the benefits derived from interpersonal agency, the Catholic Church advocates for the principle of *subsidiarity*, which states that individuals and small groups should be allowed the freedom to act and higher levels of

[33] The next few paragraphs rely on Russell Hittinger, "The Coherence of the Four Basic Principles of Catholic Social Doctrine: An Interpretation," *Nova et Vetera* 7, no. 4 (2009): 791–838.

social order should not step in to remove that agency unless the smaller levels need help.[34] Decentralized action is both more effective in many cases and more appropriate to the kinds of communal agents humans are. Yet, due to human finitude and sinfulness, sometimes these lower levels of social organization will either deny some people the option or necessary resources and tools to participate or will lack the means to provide them. In such situations, higher levels of social organization like the state must step in to support and defend the rights of individuals to allow them to more fully participate in society.[35] Agency is thus always oriented toward the communal.

This suggests why there is a close tie between individual agency and Catholic social teaching (CST). The capacity for responsibly acting for our own futures is among the foundational principles of CST. As Pope Leo XIII wrote in 1891,

> For man, fathoming by his faculty of reason matters without number, linking the future with the present, and being master of his own acts, guides his ways under the eternal law and the power of God, whose providence governs all things. Wherefore, it is in his power to exercise his choice not only as to matters that regard his present welfare, but also about those which he deems may be for his advantage in time yet to come.[36]

This notion of responsible human agency undergirds both the right to property and the right to a just wage in order to provide for oneself and one's family. *Quadragesimo Anno* (1931) expands this idea of the worker's

---

[34] Pontifical Council for Justice and Peace, *Compendium of the Social Doctrine of the Church*, Libreria Editrice Vaticana & USCCB, 2007, 81–82, with key ideas from Pope Pius XI, *Quadragesimo Anno*, 1931, §79, vatican.va/content/pius-xi/en/encyclicals/documents/hf_p-xi_enc_19310515_quadragesimo-anno.html. See also John XXIII, *Mater et Magistra*, 1961, §117, vatican.va/content/john-xxiii/en/encyclicals/documents/hf_j-xxiii_enc_15051961_mater.html.

[35] David Hollenbach, *The Common Good and Christian Ethics* (Cambridge University Press, 2002).

[36] Leo XIII, *Rerum Novarum*, §7.

participation in the management of industry through its ideas of a corporate order, including entities like cooperative businesses as well as an affirmation of subsidiarity.[37] *Mater et Magistra* (1961) affirmed the increasing intensity of social relationships arising as part of the postwar state that protected the rights of people through public support.[38] Thus, as we will show in more detail in chapter 9, CST affirms both the importance of individual agency as well as its ultimately social nature.

The most fundamental relational aspect of our free agency is relationship with the Creator: God wills for us to act freely because God desires for us to freely love God, as God freely loves us. Thus, union with the will of God brings the ultimate freedom, as it is through relationship with God that we find the fulfillment of the very purpose of our existence. Christ Himself, who perfectly unites the human with the divine will, is both the Exemplar and Grace through whom we can seek this free, loving relationship with God. As exhibited in the life of Mary, true human freedom involves complete reliance on grace and decisive acceptance of God's call. The ultimate test for any technological ethical analysis is whether we are careful stewards of our trust and attention such that we allow grace above all else to lead our acts—not instinct, mere preference, fleeting pleasure, nor the persuasion or power of another.

Responsible Agency and Virtue

Because human agency is ordered toward the ultimate good, people can be judged as to whether our actions actually fulfill this end. Other people, and especially God, can judge whether our actions meet the standard of the good. Our actions are thus responsible to others, to God, and to our own purpose. We are responsible for ourselves and those entrusted to us either by the traditional duties we have to those closest to us or by the exigencies of need, exemplified by the case of the Good Samaritan (Luke 10:25–37).

[37] Pius XI, *Quadragesimo Anno*, §80, 91–96.
[38] John XXIII, *Mater et Magistra*, §59–67.

Acting responsibly requires virtues, character traits that incline us toward choosing the good. Virtues are necessary for human flourishing because human happiness ultimately lies in the kind of person each of us is, rather than anything external that any individual possesses. This is the primacy of being over having.[39] Chief among the natural virtues is prudence or practical reason, which is the standing disposition to act well by effectively pursuing real goods in line with the moral law while assessing and avoiding potential dangers. To seek real ends, people need their other affective capacities to align with the good so that they are not misled. They need temperance, so they are not overwhelmed by temptations of pleasure; justice, so they are not tempted by greed; and courage, so that fear of danger does not turn them from the good. These virtues are all necessary for responsible agency. In turn, growing into these virtues itself requires the exercise of responsible agency, since virtues are habits that only arise through the continued process of choosing well. Because of the causal loop between responsible agency and virtue, individuals must be supported by a cultural milieu and by social practices that enable and reward responsible agency.

In conclusion, people are made to act well with others. We "image" the creative and loving power of God by effectively stewarding our world and serving others. Yet this is not only a duty. Responsible agency is also how people realize joy and fulfillment. Whether it be fixing a car, coding a computer program, mentoring another in developing a skill, or coming to another's aid through charitable action, we pursue these ends in cooperation with others. It is only through such relational agency that we flourish. Yet, social influences and conditions can also impede our agency and flourishing. The next chapter will turn to the positive and negative effects of others on individual agency.

[39] John Paul II, *Centesimus Annus*, §36, vatican.va/content/john-paul-ii/en/encyclicals/documents/hf_jp-ii_enc_01051991_centesimus-annus.html.

Chapter 3

Human Agency as Conditioned

As the last chapter described, human agency arises out of our nature as rational beings created in the image of God. Our intellect and will allow us to freely seek the good. Yet our agency is not unconditioned. It is neither absolute nor arbitrary. Human agency encounters boundaries that arise from the inherent limitations of our nature, from the laws of the physical world, from the demands of the intrinsic values of the created world, from the societal rules designed to organize collective existence, and from other constraints imposed by living with others. These limits are not purely negative. In fact, our relationships to others and to reality more broadly enable productive agency. Yet, because of the fallen, sinful state of the world, these conditioning factors often corrode and corrupt agency. Because we are relational beings, virtuous agency depends in many ways on others, is reflected in our interactions with others, and is demanded in light of the existence of the other.

Positive Conditions for Human Agency

Responsible agency can be enabled by others, and it is brought to perfection through collaborative activity. Our "others" include God foremost of all, who enables freedom by bringing us into existence, confronting us with reality, and granting us the grace to exceed our human capacities for happiness. They also include other people, both near relations as well as all who share our social life. They even include preceding generations, who developed the traditions and technologies

through which we engage the world. We therefore cannot truly understand agency without understanding the contributions of others.

Human Agency as Created

Constraints shape the contours of human freedom because "man is not only a freedom which he creates for himself. Man does not create himself. He is spirit and will, but also nature."[1] We have not brought ourselves into existence, nor have we created the universe. Such is the work of God's primary causality, meaning that God is the ultimate cause of all existence, in a manner that is both immediate and ongoing.[2] Yet, our status as creatures does not reduce humans to passivity. Although creation receives its existence from God's creative power, God generously grants created beings genuine causal efficacy.[3]

This causality in creation is what is called "secondary causality," which reaches its highest form in the agency of rational beings. In this created order, "God willed to leave man in the power of his own counsel, so that he would seek his Creator of his own accord and would freely arrive at full and blessed perfection by cleaving to God."[4] Some might think that an all-powerful God would mean that there is no power left for humanity, that, in a zero-sum game, God's strength means our weakness. But instead, God's power empowers us. God makes our strength possible. Since our agency originates in the Creator, we find in God not a rival but the guarantor and sustainer of human freedom: "God's plan poses no threat to man's genuine freedom; on the contrary, the acceptance of God's plan is the only way to affirm that freedom."[5] Consequently, the relationship between divine causality and human agency is not competitive but

---

[1] Benedict XVI, "*Address to the Bundestag*, Berlin (September 22, 2011)," AAS 103 (2011): 664.

[2] Aquinas, *ST* I, qq. 44 and 45.

[3] Aquinas, *ST* I, qq. 104 and 105.

[4] Second Vatican Council, *Gaudium et Spes*, §17.

[5] John Paul II, *Veritatis Splendor*, 1993, §45, vatican.va/content/john-paul-ii/en/encyclicals/documents/hf_jp-ii_enc_06081993_veritatis-splendor.html.

complementary. The illusion that true agency requires total, unfettered, and absolute freedom overlooks a crucial reality: A freedom that is not rooted in truth turns against man and ends by destroying itself.[6] A freedom that rejects God will inevitably undermine itself by turning away from its ultimate fulfillment.

Human freedom, then, possesses an "essential and constitutive relationship to truth."[7] Action never occurs in isolation; many constraints are not arbitrary but flow from God's design, which wrote into the world "an order and a dynamism that human beings have no right to ignore."[8] In the attentive consideration and respect of our own being and of surrounding reality, we find a structure within which to navigate the world effectively. Consequently, rather than being mere hindrances, limitations reveal the truth of our surroundings by exposing their inner workings and potential. Our very needs and limitations also provide opportunities for free acts of love by others. Boundaries invite us to recognize the reality of others, making space for them in our choices and actions.

Through engagement with reality, we continually expand the scope of human agency and make new types of actions possible. This is particularly evident in our use of technology, which may expand or limit our capabilities in various ways. Technologies grant us vastly increased power over the surrounding world, but, as chapter 6 discusses, they can also lead us to neglect skills and thereby decrease our abilities. Moreover, technologies can expand inequality and be used by those in power to oppress others or destroy the environment upon which we depend. As Pope Francis argued, "We have the freedom needed to limit and direct technology; we can put it at the service of another type of progress, one which is healthier, more human, more social, more integral."[9] This other

[6] Congregation for the Doctrine of the Faith, Instruction on Christian Freedom and Liberation: "The Truth Makes Us Free," March 22, 1986, §26.
[7] John Paul II, *Veritatis Splendor*, §4.
[8] Francis, *Laudato Si'*, §221.
[9] Francis, *Laudato Si'*, §112.

type of progress does not arise from external objects, commodities, or technology, but from the cultivation of personal freedom, nurtured and allowed to grow rather than left to erode.

Growth in freedom is not merely an infinite capacity for making choices but also must involve growth in self-possession. Subordinating our freedom to external objects alienates us and robs us of true autonomy. It does not lead to being "more human, more social, more integral." Yet, growth in human freedom is also a growth in the giving of oneself, because possessing oneself without the intention of giving and receiving does not lead to progress or flourishing, but to solitude, stagnation, and the isolation of a will curved in upon itself. Our likeness to God "reveals that man, who is the only creature on earth which God willed for itself, cannot fully find himself except through a sincere gift of himself."[10] This "sincere gift of self . . . is what gives the life and freedom of the person their truest meaning."[11] It is in relation to others that individual freedom becomes meaningful and real.[12]

More fundamentally, agency requires the presence of our Creator, in whom we find not only the origin of our freedom as a gift from God, but also its destiny, as St. Paul reminds us: "For you were called to freedom" (Galatians 5:13). This vocation to live in freedom is made possible by yet another harmonious collaboration "between the Lord's grace and human freedom, between gift and task."[13] Alongside the gift of freedom, God bestows grace, which not only elevates human agency beyond the natural order but also orients it toward communion with each of the Trinitarian Persons. Through the supernatural aid of grace, human persons are invited to participate in the Trinitarian life, where we share in God's divine existence. Consequently, the growth of our agency during our journey on earth is "made possible by grace, which enables us to possess the full

[10] Second Vatican Council, *Gaudium et Spes*, §24.
[11] John Paul II, *Evangelium Vitae*, §96.
[12] See also Arendt, *The Human Condition*, 46.
[13] John Paul II, *Veritatis Splendor*, §24.

freedom of the children of God (cf. Romans 8:21) and thus to live our moral life in a way worthy of our sublime vocation as 'sons in the Son'."[14] We cannot reach ultimate happiness by our own efforts; instead we receive it as a gift.

Human beings are not left alone to navigate an uncertain future, wondering whether we can fulfill our calling. It is within this relationship with the mystery of the Triune God—who generously shares divine freedom—that our own freedom finds its true fulfillment. As human beings, we always stand before "the spiritual horizon of hope, thanks to the *help of divine grace* and with *the cooperation of human freedom.*"[15] In this divine-human collaboration, we discover that true freedom is not limitless autonomy but a participation in God's own triune life.

Human Agency and Sociality

Our lives, and thus our agency, are not only a gift from God, but also fundamentally a gift from others who came before and now surround us. For example, when Aquinas describes the virtue of piety, he does not merely look at our relationship with God. Instead, he describes piety as a virtue of gratitude toward all sources of our existence, or "principles of our being and government."[16] Those sources include parents and country, along with kin and fellow citizens. In this recognition of the importance of others to our lives, he follows the insights of ancient philosophers. Socrates refused to evade his death sentence because he felt that obedience to the laws responsible for his own formation was warranted,[17] and Cicero includes piety toward country and parents as part of the natural law.[18] It is imperative to respect those who have helped shape us, a responsibility that

[14] John Paul II, *Veritatis Splendor*, §18.
[15] John Paul II, *Veritatis Splendor*, §103, italics original.
[16] Aquinas, *ST* II-II, a. 101, a. 1.
[17] Plato, "Crito," in *Complete Works*, ed. John Cooper, trans. G. M. A. Gruber (Hackett, 1997), 37–48; 50a–54e.
[18] Cicero, *Cicero: On Invention. The Best Kind of Orator. Topics. A. Rhetorical Treatises*, trans. H. M. Hubbell (Harvard University Press, 1949), II.22.

undergirds the commandment to honor father and mother (Exodus 20:12; Deuteronomy 5:16). Others shape not only our existence but also how we use our reason and free will.

Most obviously, this shaping occurs through education. In this discussion of education, we include the broadest effects of human formation, rather than just the acquisition of knowledge, since Catholic education seeks to form the whole person aimed at wisdom. Through the encouragement, discipline, and instruction of others, we learn how to act, think, and even feel.[19] This formation occurs most fundamentally in the family, which bears the first responsibility for education. Each family has its own culture that shapes the child, and every parent must learn how to foster a child's particular talents and vitality while also restraining their incipient vices. This early formation sets the stage for all future action.

Next, a child is shaped by many years of formal education, both through instruction in content and in the skills of reasoning, but also through the hidden curricula structuring how we act in society. Again, children learn reason, but they also have their dispositions shaped in a way ordered toward right action both explicitly and implicitly. Yet formation extends beyond those people specifically assigned responsibility for it, like teachers and parents, and includes the informal encounters of everyday life. Every interaction with people in the community or parish shapes the child, media forms thought, and laws set the structure for action. Our agency cannot come into existence save for its formation by family, teachers, and community.

At a deeper level, our very media and tools of thought depend on those who came before us. Language is a foundational medium of rational thought.[20] In this, humans reflect Christ, the divine Word. In English, this description of Christ translates the Greek word *Logos*, which has a range of

[19] For a discussion of the importance of the formation of sentiment by education, see C. S. Lewis, *The Abolition of Man* (Simon & Schuster, 1978).

[20] For a recent elaboration of this idea, shared by many scholars, see Charles Taylor, *The Language Animal: The Full Shape of the Human Linguistic Capacity* (Harvard University Press, 2016).

meaning, including reason, speech, and language. No one develops their own language. Though language is essential to living as a rational agent, it is something each person receives from prior generations. Similarly, intellectual and cultural traditions set the terms for our thoughts and questions.[21] Everything we think about and all the sorts of action we consider taking are shaped by a long line of people passing down ideas. We are part of a chain of history.

It is for these reasons that Charles Taylor argued that appropriating individual agency cannot refer to some kind of radical individualism or pure self-assertion.[22] Instead, it means accepting our historical situatedness, recognizing the role of the communities, family, and tradition that have shaped us. To live well as agents, we must come to accept these effects with gratitude, even though sometimes acting well requires criticizing times when these traditions and sources do not live up to their own best intuitions or fall away from truth. That is the topic of the next section. But to put the insight of this section simply, a person cannot be an agent alone. Our existence is received from and supported by God, and our God-given agency is shaped by multiple sources.

Negative Influences on Human Agency

Insofar as our human agency is shaped by other people who surround us, whether the nearest community of the family or the many others who form us, our development as persons and as moral agents is conditioned by our social contexts. Yet all human societies are characterized by both individual and structural sin. Thus, we are not only formed as persons in society but also, inevitably, malformed by the moral failings of the societies in which we live.

The Throwaway Culture and the Technocratic Paradigm

[21] Alasdair MacIntyre, *Whose Justice? Which Rationality?* (University of Notre Dame Press, 1988).

[22] Charles Taylor, *The Ethics of Authenticity* (Harvard University Press, 2018).

Social contexts condition our formation as moral agents in two intertwined ways. First, the societies in which we live cultivate not only particular virtues but also particular vices in the human person. Human moral formation cannot be reduced to these broader forces, yet our sinful societies enact a seductive pedagogy to which none of us are immune, shaping us through what the Church calls "structures of sin."[23] Second, the societies in which we live present us with particular conditions for exercising our agency. The social structures in which we are embedded often limit the range of available actions. Just as significantly, these structures often distort our ethical perception and limit our recognition of the options for action. The actions that may seem obvious, possible, and just from the outside are constrained by the structures of the societies in which we live. Because our actions shape who we are as persons, these constrained conditions for exercising agency also tend toward the cultivation of particular moral qualities.[24] As a result, these two modes of moral (mal)formation reinforce one another.

While all human societies are characterized by both individual and structural sin, particular societies exhibit and cultivate their own particular moral distortions. For our purposes it is particularly relevant to examine how the pervasive social logic of the "technocratic paradigm" distorts contemporary societies.[25] As described in *Laudato Si'* §101, the "dominant technocratic paradigm" provides a means of analyzing "the place of human beings and of human action in the world." The result of the technocratic paradigm is "to make the method and aims of science and technology an

[23] John Paul II, *Sollicitudo Rei Socialis*, December 30, 1987, §36–40, vatican.va/content/john-paul-ii/en/encyclicals/documents/hf_jp-ii_enc_30121987_sollicitudo-rei-socialis.html; John Paul II, *Evangelium Vitae*, §11–24.

[24] See Daniel J. Daly, *The Structures of Virtue and Vice* (Georgetown University Press, 2021); Daniel K. Finn, ed., *Moral Agency Within Social Structures and Culture: A Primer on Critical Realism for Christian Ethics* (Georgetown University Press, 2020); Daniel K. Finn, "What Is a Sinful Social Structure?" *Theological Studies* 77, no. 1 (2016): 136–164.

[25] We provide an extended analysis of this concept in AI Research Group, *Encountering Artificial Intelligence,* 6–8.

epistemological paradigm which shapes the lives of individuals and the workings of society" (§107) and thus privilege technological power and "progress" over any other consideration, including the flourishing of human lives.

The result of this paradigm is a reductionist understanding of the human person, "dominated by [the] internal logic" (§108) of a technocratic paradigm that "shapes the lives of individuals and the workings of society" (§107). This understanding of the individual person, of human society, and of the created order shapes us as moral agents, coaxing us to embrace the utilitarian logic that "accepts every advance in technology with a view to profit, without concern for its potentially negative impact on human beings" (§109). The power, wealth, and comfort promised by modern technologies, whether AI tools or other innovations, condition us to see these technologies and their results as fundamentally good. Yet the result of the technocratic paradigm is what Pope Francis names as a "throwaway culture" that says to other people: "I use you insofar as I need you. When I am not interested in you any more, or you are in my way, I throw you out."[26]

We might say, then, that this is a vicious pedagogy that encourages us to embrace greed and pride while rejecting the virtue of charity. Just as significantly, the technocratic paradigm generates a "fragmentation of knowledge" that "makes it difficult to see the larger picture" (§110). It is not simply that this set of social structures guides us away from virtue and toward vice, but also that it distorts our perception of the world, "lead[ing] to a loss of appreciation for the whole, for the relationships between things, and for the broader horizon," and thereby "mak[ing] it hard to find adequate ways of solving the more complex problems of today's world"

[26] Francis, *Angelus*, January 29, 2023, vatican.va/content/francesco/en/angelus/2023/documents/20230129-angelus.html. Cf. *Laudato Si'*, § 20–22.

(§110).[27] This limitation in perspective further constrains and conditions our actions, reinforcing patterns of vice.

Economic Structures

The throwaway culture and the technocratic paradigm can be found in aspects of current socioeconomic structures and the dynamics that they foster locally, regionally, and globally. As we will discuss in Part II, businesses are developing new technologies of control and surveillance that increasingly threaten moral agency and exploit democratic processes. Technological advances like AI can enable the manipulation of workers, consumers, companies, and societies. Such technologies demand a critical assessment informed by a vision of a just society.[28] This vision strives to address, contain, and hopefully remove inequities that inhibit human and social flourishing in order to promote social, cultural, and economic development.[29]

One of the primary ways that economic structures can negatively affect agency is by denying the poor access to the resources necessary to act well with relative ease. For example, the accumulation of very high profits in the hands of a few investors combined with lack of social investment leads to severe inequality. This inhibits the increased societal participation, and thus agency, of the marginalized by denying them education, political participation, and the opportunity to choose good work. This pattern of accumulation, seen in the massive profits of many technology companies, thus serves the logic of the technocratic paradigm rather than justice.[30] To take a second example widespread in the nineteenth through the twenty-

---

[27] Cf. Francis, *Laudato Si'*, §20: "Technology . . . proves incapable of seeing the mysterious network of relations between things and so sometimes solves one problem only to create others."

[28] See Kenneth R. Himes, "Catholic Social Teaching on Building a Just Society: The Need for a Ceiling and a Floor," *Religions* 8, no. 4 (2017), doi.org/10.3390/rel8040049.

[29] See Francis, *Laudato Si'*.

[30] Cf. Karen Lebacqz and Matthew J. Gaudet, *Eight Theories of Justice: Perspectives from Philosophical and Theological Ethics* (Fortress Press, 2025), 111–117.

first centuries, extractive logics, which characterize colonial processes, exacerbate social injustices, maintain workers in inhumane working conditions, stifle economic growth, and greatly harm the environment.[31]

What we think is important to highlight here is the effect of current structures on work, one of the fundamental arenas where human stewardship and agency are enacted. Today, workers face grave dangers to their quality of life and working conditions, with many forms of dehumanizing work. In *Gaudium et Spes*, the bishops of the Second Vatican Council listed "disgraceful working conditions" among the contemporary threats to human dignity.[32] John Paul II's encyclical letter *Laborem Exercens*, which focuses primarily on human work, highlighted the preeminence of the worker as the subject, not the object, of work. John Paul II warned against the commodification of persons and indicated the dignity of work as a cooperation in God's continuing work of creation.[33] Work *should* allow human beings to more deeply realize their human identity as made in the image and likeness of God, but it often does not.

---

[31] See Macarena Gómez-Barris, *The Extractive Zone: Social Ecologies and Decolonial Perspectives* (Duke University Press, 2017); Peter Hughes, SSC, "The Pan Amazon, Extractive Industries, and the Church," in *Fragile World: Ecology and the Church*, ed. W. T. Cavanaugh (Wipf & Stock, 2018), 97–111; Henry Veltmeyer and Arturo Ezquerro-Cañete, eds., *From Extractivism to Sustainability: Scenarios and Lessons from Latin America* (London: Routledge, 2023); Caesar A. Montevecchio and Gerard F. Powers, eds., *Catholic Peacebuilding and Mining: Integral Peace, Development, and Ecology* (Routledge, 2022); Simone M. Müller, Matthias Schmidt, and Kirsten Twelbeck, eds., *Ecological Ambivalence, Complexity, and Change: Perspectives from the Environmental Humanities* (Routledge, 2025); Kate Crawford, *Atlas of AI*.

[32] For the Council Fathers, "whatever insults human dignity, such as subhuman living conditions, arbitrary imprisonment, deportation, slavery, prostitution, the selling of women and children; as well as disgraceful working conditions, where men are treated as mere tools for profit, rather than as free and responsible persons; all these things and others of their like are infamies indeed." See Second Vatican Council, *Gaudium et Spes*, §27.

[33] See John Paul II, *Laborem Exercens*, 1981, https://www.vatican.va/content/john-paul-ii/en/encyclicals/documents/hf_jp-ii_enc_14091981_laborem-exercens.html. See also Patricia A. Lamoureux, "*Laborem Exercens*," in *Modern Catholic Social Teaching: Commentaries and Interpretations*, ed. K. R. Himes, Lisa Sowle Cahill, Charles E. Curran, David Hollenbach, and Thomas Shannon (Georgetown University Press, 2018), 403–428.

The current dominant economic model—centered on production, consumption, and the pursuit of what is not essential—appears to be characterized by processes that affect human beings, perverting individual and collective working experiences, compromising the experience of work, and increasing social inequities.[34] Among the competing models that have been proposed, we will highlight two. The first is Taylorism, which takes its name from the American engineer Frederick Winslow Taylor (1856–1915). This model focuses on industrial management, workflows, and work processes. The multiple work steps in a production process are analyzed, optimized, standardized, and placed in a predetermined sequence. The goal is to expand efficiency and productivity, reduce costs, and increase the quality of the products.[35] This approach focuses on productive systems. However, it can neglect the psychological and social dimensions of work, as disenfranchised workers come to feel like cogs in a machine. Concerns with this model will be discussed below in a section on deskilling.

In his encyclical letter *Caritas in Veritate,* Benedict XVI argued for an alternative systemic and structural model: an economy of communion.[36] For Benedict, the goal is to avoid polarization between for-profit companies and nonprofit organizations. He envisioned "traditional companies which nonetheless subscribe to social aid agreements in support of underdeveloped countries, charitable foundations associated

[34] See Daniel K. Finn, *Consumer Ethics in a Global Economy: How Buying Here Causes Injustice There* (Georgetown University Press, 2019).

[35] See Frederick Winslow Taylor, *The Principles of Scientific Management* (Norton, 1967).

[36] See Meghan J. Clark, "*Caritas in Veritate*," in Himes et al., eds., *Modern Catholic Social Teaching*, 504–507. See also Stefano Zamagni, "The Economy of Communion Project as a Challenge to Standard Economic Theory," *Revista Portuguesa de Filosofia* 70, no. 1 (2014): 44–60; Luigino Bruni and Stefano Zamagni, "The 'Economy of Communion': Inspirations and Achievements," *Finance et Bien Commun* 3, no. 20 (2004): 91–97; Luigino Bruni and Tibor Héjj, "The Economy of Communion," in *The Palgrave Handbook of Spirituality and Business*, ed. L. Bouckaert and L. Zsolnai (Palgrave Macmillan, 2011), 378–386.

with individual companies, groups of companies oriented towards social welfare."[37] This would mean

> a broad new composite reality embracing the private and public spheres, one which does not exclude profit, but instead considers it a means for achieving human and social ends. Whether such companies distribute dividends or not, whether their juridical structure corresponds to one or other of the established forms, becomes secondary in relation to their willingness to view profit as a means of achieving the goal of a more humane market and society.[38]

Such companies would better realize the communal nature of human agency.

This example suggests that structural solutions can better realize a just society. Attention to specific cultural and social contexts, particularly in the Global South, can further enrich and nuance the critical assessment of existing social structures, systems, and economic dynamics in ways that foster social justice and that rely on diverse contributions aimed at promoting the common good.[39]

Institutional Structures and Bureaucracy

Another social element of modern life that can negatively, although not only negatively, affect agency is bureaucracy. The pioneering sociologist Max Weber defined bureaucracy in terms of a hierarchical, professional system of organization governed by impersonal, calculable rules.[40] Many would add that features of quantitative analysis like metrics and cost-

---

[37] Benedict XVI, *Caritas in Veritate*: *On Integral Human Development in Charity and Truth*, 2009, vatican.va/content/benedict-xvi/en/encyclicals/documents/hf_ben-xvi_enc_20090629_caritas-in-veritate.html, §46.

[38] Benedict XVI, *Caritas in Veritate*, §46.

[39] For example, see Agbonkhianmeghe E. Orobator, "*Caritas in Veritate* and Africa's Burden of (under)Development," *Theological Studies* 71, no. 2 (2010): 320–334.

[40] Max Weber, "Bureaucracy," in *From Max Weber*, ed. H. H. Gerth and C. Wright Mills (Oxford University Press, 1958), 196–244.

benefit analysis are now central to bureaucracy, at least since the late nineteenth century.[41] This quantitative, algorithmic aspect is critically important regarding the bureaucratic use of AI.

There is a paradox to the effects of bureaucracy. Because of their size, large contemporary institutions would be impossible without a bureaucratic framework to support them. As Alasdair MacIntyre noted, no practices dedicated to a human good can survive long without the support of an institution: Medicine needs hospitals, scholarship needs universities, sports need leagues and governing bodies.[42] Institutional structures are vital arenas for shaping our virtues and agency, and bureaucracy is necessary for developing and maintaining them. Well-functioning bureaucracies can allow for a coordination that supports agency.

At the same time, bureaucracy and institutions often threaten to undermine practices and agency, forcing people to follow cumbersome rules that serve only to perpetuate the bureaucratic system or pursue ends that are not intrinsically good. Given its complexity, modern medicine could not exist without the bureaucracy of the healthcare system, yet those same institutional forces frequently stymie the actions of healthcare practitioners. Doctors and nurse practitioners are forced to seek insurance approvals or must forgo their best clinical judgment in order to follow rules and guidelines.[43] Bureaucracy is thus necessary for but also corrosive of agency.

Scholars have noted three ethical effects of bureaucracy. First, the focus on impersonal rules in bureaucracy can redirect people away from a focus

---

[41] See Theodore Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton University Press, 1996).

[42] Alasdair MacIntyre, *After Virtue: A Study in Moral Theory*, 2nd ed. (University of Notre Dame Press, 1984), chapter 14.

[43] Ronald W. Dworkin, "Doctors and the Danger of Rule-Dependency," *The American Interest* (blog), April 29, 2020, the-american-interest.com/2020/04/29/doctors-and-the-danger-of-rule-worship/; Paul Scherz, *The Ethics of Precision Medicine: The Problems of Prevention in Healthcare* (University of Notre Dame Press, 2024).

on virtue.[44] Virtue is neither impersonal nor rule-based. Instead, it depends on decisions emerging from the prudential judgment of an agent with a well-formed character. The temptation to replace virtue with rule-following becomes especially strong with the emergence of highly metrics-driven systems. The historian Ted Porter chronicles how many modern quantitative bureaucratic procedures arose from an explicit rejection of prudence as a ground for decision-making due to calls for democratic transparency.[45] Instead, these procedures focused on seemingly objective processes like cost-benefit analysis and algorithms. (We will explore this aspect of modern bureaucracy further in Part II as we examine algorithmic governance.) Bureaucracy thus shifts agency away from virtuous responsibility, entrusting responsibility to impersonal rules.

Bureaucracy undermines responsible agency in another way by disconnecting our actions from their effects. The theologian Romano Guardini describes how modern social systems have grown so large and complicated that an individual never sees the concrete effects of their decisions.[46] A person makes a choice, but the effects end up reverberating through the bureaucratic system such that it can feel as if the action had no effect. Actions seem to become agentless, impersonal. Similarly, people often have no recourse when caught in automated phone systems or become subject to decisions made by algorithms rather than agents (both in the philosophical sense and as holders of the modern job title). Guardini attributes the modern crisis of responsibility in part to this aspect of bureaucracy.

CST has criticized bureaucracy for these and other reasons. Bureaucracies are overly expensive and frequently ineffective:

---

[44] MacIntyre, *After Virtue*.

[45] Porter, *Trust in Numbers*.

[46] Romano Guardini, *The End of the Modern World* (ISI Books, 2001), 122–126. For discussion in relation to AI, see Scherz, *Tomorrow's Troubles*, chapter 8.

> International organizations might question the actual effectiveness of their bureaucratic and administrative machinery, which is often excessively costly. At times it happens that those who receive aid become subordinate to the aid-givers, and the poor serve to perpetuate expensive bureaucracies which consume an excessively high percentage of funds intended for development.[47]

These concerns lead to desires for a "more devolved and organic system of social solidarity, less bureaucratic but no less coordinated."[48] Pope Francis suggests that the technocratic methodologies inherent to contemporary bureaucracy can shape our worldview.[49] They form us in a technocratic paradigm rather than an openness to encounter, especially as their formal systems block embodied engagement with others.

Bureaucracies are already deploying AI because AI, and computers more generally, are perfectly fitted for its world of impersonal rules and quantitative metrics.[50] AI increases the bureaucratic ability to gather and analyze data and surveil a company's workers and wards. In the abstract, it is possible to argue that AI might reduce human engagement with bureaucracy. Since it is perfectly suited for bureaucratic tasks, it could remove such boring work from the human purview, allowing people to focus on the actual goods of encounter and practice, while machines handle the "back office" tasks required by bureaucracies. Experience with prior generations of information technology suggest, however, that giving bureaucracies new tools is likely to merely expand the activities of bureaucracies. As we will discuss in relation to rapidification, the volume of bureaucratic communication increased with the introduction of email. Rather than easing the work of typing and mailing letters, email generated more messages in need of response. Similarly, digital technologies allow for greater surveillance and reporting. It is therefore unlikely that AI will

[47] Benedict XVI, *Caritas in Veritate*, §47.
[48] Benedict XVI, *Caritas in Veritate*, §60.
[49] Francis, *Laudato Si'*.
[50] Neil Postman, *Technopoly: The Surrender of Culture to Technology* (Vintage, 1993), 107–121.

decrease the reach of bureaucracy. It could instead expand it, which would likely intensify its negative effects on responsible agency.

Though it is a capacity inherent to our nature, human agency is fundamentally shaped by others. Most importantly our agency is shaped by God, its source as well as the end to which it is directed. Others also form our agency through our upbringing and enable it through ongoing cooperative endeavor. We depend on others for our language and practices. Even our technologies, like AI, through which we can think and care for the world, come from others. Yet, others may also malform our agency by shaping deficient worldviews, unjust economic structures, and stifling institutional forms. Technologies can also be used to reinforce these destructive influences on our freedom. In Part II, we will explore specific examples of how AI is used to reinforce problematic effects on our agency. But first, the next chapter will discuss the question of whether AI itself could be considered an agent.

Chapter 4

Can an AI have True Agency?

Before turning to AI's specific effects on human agency, it seems important to ask whether an AI system itself could exhibit agency within the framework of human agency developed in previous chapters. Some developers and tech firms tell us that there are already AI "agents." Amazon, for example, describes an AI agent as "a software program that can interact with its environment, collect data, and use the data to perform self-determined tasks to meet a predetermined goal."[1] This description is akin to that of a "travel agent," who determines the best flights and books hotels using dates and destinations predetermined by the client. Similarly, an AI agent in this sense may choose the actions needed to achieve a goal, but the goal, and the methods used to reach that goal, are predetermined by human beings. As the last two chapters have discussed, this use of the word "agent" is different from the way human agency is understood by philosophers and theologians.

Persons vs. Things

Scholars have long explored the question of what counts as true agency through the analyses of moral concepts such as responsibility, autonomy, and personhood. For example, in the Enlightenment era, Kant famously distinguished between persons and things: "Beings, . . . , if they are beings without reason, still have only a relative worth, as means, and are therefore called things, whereas rational beings are called persons because their

[1] Amazon.com, "What Are AI Agents?" aws.amazon.com/what-is/ai-agents/.

nature already marks them out as an end itself, that is, as something that may not be used merely as a means. . . ."[2] Similarly, the Catholic philosopher Robert Spaemann, in his book on persons, differentiates "someone" from "something," saying, "Within the universe of things that exist, 'persons' have a special position."[3] Notably, Kant believed that nonhuman animals could not have their own ends. For example, when discussing the moral duties of persons, he denied that "this supposed duty can be referred to objects other than persons . . . [such as] the part of nature endowed with sensation and choice (animals)."[4] Nonhuman animals, therefore, are relegated by Kant to the "things" category.

But this is a potentially awkward classification to maintain, as even casual observation will confirm that some nonhuman animals have more in common with humans than they do with inanimate objects or machines. To be fair, Kant's rigid person/thing classification system, made in the context of morality, does not attempt to address this intuitive truth. Elsewhere, however, he does speak to the obvious difference between nonhuman animals and inanimate objects. The basic distinction he maintains is that animals, unlike rocks, for example, possess internal organization. For our purposes, this is important because it is in this context that he contrasts their internal organization with the external organization of machines: "A machine has only a motive force; an organized being possesses within itself a formative force, and this it communicates to its materials, which do not possess it of themselves. Thus,

[2] Immanuel Kant, *Groundwork for the Metaphysics of Morals*, trans. and ed. Mary Gregor, rev. ed. (Cambridge University Press, 2012), Ak. 4:428.

[3] Robert Spaemann, *Persons: The Difference Between 'Someone' and 'Something,'* trans. Oliver O'Donovan, repr. ed. (Oxford University Press, 2017), 9.

[4] Immanuel Kant, *The Metaphysics of Morals*, trans. and ed. Mary Gregor (Cambridge University Press, 1996), Ak. 6:442.

it organizes them."[5] In the Aristotelian tradition, this formative organization is termed the soul.[6]

Despite the obvious differences between nonhuman animals and machines, however, Kant believes that neither kind of organization, whether the internal, organic formative force of animals, or the external, mechanistic, motive force of machines, is useful in explaining the kind of agency that humans enjoy. Kant insists that the "internal organization" that nonhuman animals share with humans is not relevant for morality, because nonhuman animals only act from the instincts that they find themselves already possessing, rather than laws that they give themselves. In the case of machines, the ends are ultimately given by human intelligence; in the case of nonhuman animals, the ends are ultimately given by nature. Thus, only persons have autonomy in the most robust sense of ability to decipher and freely choose to adhere to laws.

This distinction between the autonomy of humans and the autonomy of nonhumans is also explored in science fiction. The 2023 movie *Simulant* imagines a future where humans live together with human-like AIs that appear so much like humans that special detection equipment is needed to tell who is human and who is AI. The most important distinction is that the simulants are programmed to comply with a version of Isaac Asimov's well-known "Three Laws of Robotics."[7] So, while a typical human willingly follows laws such as "do not harm humans," a simulant is programmed, rigidly, to *never* harm humans. In the latter case, the simulants are following a law that was given to them; in Kantian terms, they are "acting according to laws." In the former case, the humans give themselves a law, or, in Kantian terms, they are "acting according to principles." The movie further imagines human rebels who are determined

[5] Immanuel Kant, *Critique of Judgment*, trans. Paul Guyer and Eric Matthews (Cambridge University Press, 2000), Ak. 5:374.

[6] Aristotle discusses the soul in many works, but especially in his *De anima*.

[7] April Mullen, dir., *Simulant* (WANGO Films, Myriad Pictures, Particular Crowd, 2023), 1h35m; Isaac Asimov, "Runaround," in *I, Robot* (Random House Worlds, 2004), 25–46.

to remove the laws from these simulants, convinced that they, like humans, should be "free" and "have true autonomy."

These distinctions, explored both by moral philosophy and by science fiction, help answer the question of what it might mean for a system to have genuine autonomy but not the same kind of autonomy as humans have. We can say that humans are autonomous in the sense that we pursue means to achieve certain ends but also in the sense that we deliberate about the ends themselves. It is possible to believe that humans are "programmed" by nature to pursue the ends of survival and procreation, such that everything we do is ultimately directed toward that fixed end.[8] But a Catholic understanding of "person" agrees with Kant on this particular point, positing that humans do not act only according to laws that are given to us but also according to laws that we give ourselves. We create art, lay down our lives for our friends, adopt children, and go on hunger strikes—all ends that do not have a meaningful connection to the survival and reproduction of an individual organism. As discussed in chapter 2 and later in this chapter, humans are directed toward an ultimate good, but we deliberate regarding the choice of all the particular goods that serve as means to that more universal good.

Artificially intelligent systems, to the extent that they act autonomously, do not have agency in the robust sense. When an autonomous car veers onto a side road, or an LLM gives advice, those systems are acting according to their programming, even when the complexity of their actions makes it difficult to trace a specific action to a specific programming element or data set. Thus, agency in an AI is limited, bounded by the value function for which they are optimized and constrained by the algorithms, data, and fine-tuning that constitute their training, all of which is determined by their developers. In this way, AIs are similar to what Aquinas calls inanimate bodies of nature, which function "always, or nearly always, in the same way, so as to obtain the best result.

[8] Indeed, Thomas Aquinas mentions these two ends, among others, as tied to "inclinations." See Aquinas, *ST* I-II, 94.2.

Hence it is plain that not fortuitously, but designedly [*ex intentione*], do they achieve their end."[9] The AI cannot choose to do a different task nor refuse a task it has been programmed to perform. The ends are never at stake, only the means.

The Role of Interiority

AI agency also differs from human agency in a deeper sense: An AI lacks conscious interiority.[10] In other words, "it knows not what it does." Human actions reflect our emotional, psychological, and spiritual lives. While the AI might act in what looks to be a deliberate and intentional way, it lacks the "object-directed quality of human mental states—that words or actions are about something,"[11] not just as interpreted by someone else but as understood internally, by the actors themselves. John Searle examines this in his famous "Chinese Room" thought experiment, through which he claims that intentionality does not exist without a subjectively phenomenal grasp of something as meaningful.[12]

Some will argue that, while these limitations might be true of computers as we know them today, larger data sets and more and faster computing power might someday, perhaps soon, spark a qualitative leap

---

[9] Aquinas, *ST* I, q. 2, a. 3.

[10] This section draws on Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §30–35; AI Research Group, *Encountering Artificial Intelligence*, 43–105.

[11] AI Research Group, *Encountering Artificial Intelligence*, 66.

[12] Searle posits a sealed room in which there is a person with an extensive encyclopedia of Chinese words and phrases. Someone outside submits a question in Chinese and the person, who cannot read Chinese characters, simply looks up the appropriate characters to return as an answer. To the person outside, it will look as if someone in the room understands Chinese, though clearly the person inside does not. See John R. Searle, "Minds, Brains, and Programs," *The Behavioral and Brain Sciences* 3 (1980): 417–457; see also Searle's other works: "Consciousness, Unconsciousness and Intentionality," *Philosophical Issues* 1 (1991): 45, doi.org/10.2307/1522923; "Who is Computing with the Brain?" *Behavioral and Brain Sciences* 13, no. 4 (December 1990): 632–642; and *The Mystery of Consciousness* (The New York Review of Books, 1997).

toward machine consciousness and the creation of an artificial general intelligence (AGI) that is truly conscious of what it is doing and exhibits capacities of thought and action equal to or surpassing our own. Such a leap, however, is unlikely to occur through scaling current models. As Gary Marcus explains, the techniques used in LLMs like the currently best-known version, ChatGPT, "lack ways of representing causal relationships (such as between diseases and their symptoms), and are likely to face challenges in acquiring abstract ideas. . . . They have no obvious ways of performing logical inferences, and they are also still a long way from integrating abstract knowledge."[13] Current AI paradigms have no clear way to move from statistical manipulation of data to the intentional grasp of ideas.

Despite their amazing powers, today's LLMs use statistical probabilities gleaned from training on large data sets to predict a likely answer or action. The world, however, is an exceedingly complex and dynamic environment, in which unpredictable changes and exceptions abound. Inductive reasoning often breaks down in the face of the unfamiliar, for it only sees what *has been*. When confronted with experience we hypothesize, we imagine what *may* be, and we create a causal model of the world. As Erik Larson describes, this form of reasoning is conjectural, or what C. S. Peirce called abductive; we form mental models and contextual frameworks within which we evaluate and come to some understanding of new data, leaping from data to broader generalizations. New experiences are not, then, merely compared with previous data but contextualized within one or more of these mental models. Larson notes that, without abduction, computers will never "understand" anything and, thus, "there is no way for current AI to 'evolve' general intelligence . . . absent a fundamental discovery" of a way to program abductive reasoning.[14]

---

[13] Gary Marcus, *Taming Silicon Valley: How We Can Ensure That AI Works for Us* (MIT Press, 2024), 45.

[14] Erik Larson, *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do* (Harvard University Press, 2021), 274.

To build a true AGI, we would need systems that "can represent the core frameworks of human knowledge: time, space, causality, basic knowledge of physical objects and their interactions, and basic knowledge of humans and their interactions."[15] We develop these frameworks through physical and relational experience, especially in childhood. As *Antiqua et Nova* insists, the human person is a unity of "spirit and matter" who engages in "relationships with God and others" within and through an embodied existence.[16] Lacking such an existence, or, more fundamentally, a conscious experience of the world, embedding all the necessary frameworks of human knowledge in an AI is a tall order. Until we find a way to do so, machine agency will remain bounded by human history and programming methods, and it will remain subservient to our goals—which is, perhaps, a good thing.

Mistaken Interpretations of Human Agency Reflected in AI

Despite good reasons to doubt the possibility that AI could possess human agency, the idea that an AI *could* attain human agency remains persistent, both among scholars and in popular approaches to AI. This is not accidental. The idea that AI could attain human agency grows out of dominant ways of understanding the human person, and, therefore, of understanding what kind of entities could count as agents in the first place. Our imaginations and fantasies about AI are derived from existing theological assumptions about human personhood. These ways of understanding the human run contrary to Catholic anthropology, as well as other robust anthropologies, yet are pervasive in twenty-first-century culture. Because they are so pervasive, they might well be implicitly endorsed by those who, on further reflection, would reject them. It is

[15] Marcus, *Taming Silicon Valley*, 166.

[16] Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §16.

crucial, then, that we bring these ideas into the open, so we can scrutinize their implications.

Consider first *materialism*, the idea that foundationally, everything is material—not only cars, chairs, ferns, and human bodies but also consciousness, the experience of free choice, and the love a parent feels for a child. For the materialist, *matter* is all that exists. A materialist outlook is not only implicitly accepted in many of our everyday ways of navigating the world (our worlds have become, as Weber might put it, disenchanted),[17] but often explicitly endorsed in scholarly circles.[18]

When paired with questions about AI agency, materialism blurs the line between humans and machines. After all, from the perspective of many anti-materialist anthropologies, Catholic anthropology included, one difference between human agency and the activities of an AI is obvious: While the activities of an AI are completely material, human agency exceeds the material. As discussed in chapter 2, materialists cannot differentiate humans from machines in this way. For the materialist, after all, both humans and machines are no more than the stuff out of which they are composed.

Another influential theory is *behaviorism*. Behaviorism claims psychology "is the science of behavior" rather than the study of inner mental life, and that "behavior can be described and explained without making ultimate reference to . . . internal psychological processes."[19] For the behaviorist, in accounting for our mental lives—the fact and experience of human agency included—we need not appeal to anything internal. We can, instead, look exclusively to what humans *do*.

---

[17] Max Weber, "Science as Vocation," in *From Max Weber: Essays in Sociology*, ed. H. H. Gerth and C. Wright Mills (Oxford University Press, 1946), 148.

[18] A comprehensive argument for why purely material things cannot be conscious can be found in Anselm Ramelow, "In A.I., Mind Does not Matter," *Euntes Docete* 77 (2024): 27–59.

[19] George Graham, "Behaviorism," *The Stanford Encyclopedia of Philosophy,* ed. Edward N. Zalta and Uri Nodelman (Spring 2023 Edition), plato.stanford.edu/archives/spr2023/entries/behaviorism/. For the impacts of behaviorism on modern ideas of agency, see Scherz, *Tomorrow's Troubles*, 121–136.

It can sometimes be helpful to focus on human behavior while bracketing analysis of inner experiences. However, when paired with discussions of AI agency, especially when conjoined with a materialist view, a behaviorist outlook can further confuse analysis. After all, an AI can often behave in ways that mirror human behavior. The difference is in the interior life of the human being, a life of intentionality, will, and choice. For the behaviorist, this obvious difference between human agency and the operation of an AI is set aside.

Another highly influential way of understanding the human mind is *functionalism*.[20] For the functionalist, human mental life should be characterized in terms of the role it plays. If you want to understand the human mind, according to functionalists, you should pay attention to what the human mind does. That may sound like behaviorism, but there is a crucial difference. For the functionalist, we need not pay attention to the embodied behavior of human beings. Instead, we need only characterize the mind's role in abstract ways, for example, by examining what kinds of inputs lead to what kinds of outputs.

Functionalist and behaviorist thinking is in the background of the infamous Turing Test, named for Alan Turing (1912–54), one of the fathers of computer science. In his "test," Turing proposed that if a machine can imitate human communication sufficiently, then it may as well be thinking.[21] Or, to put it differently, if an AI can perform a human function, it may as well be human in that regard. With functionalism, then, the line between AI activity and human agency becomes blurry indeed.

Things do not get clearer when we layer in another influential idea: *determinism*. In its loosest sense, determinism simply states that all events

[20] See, e.g., Janet Levin, "Functionalism," *The Stanford Encyclopedia of Philosophy,* ed. Edward N. Zalta and Uri Nodelman (Summer 2023 Edition), plato.stanford.edu/archives/sum2023/entries/functionalism/. See Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §10–13.

[21] Turing, "Computing Machinery and Intelligence."

are determined. There are ways of understanding determinism that are consistent with a Catholic outlook, as chapter 3 discussed. After all, if God causes all things and stands outside of time, there is a sense in which all events are currently being caused, and therefore determined, by God. Yet, because God transcends creation, created beings maintain their own free causality. In contemporary contexts, however, that is not how determinism is typically understood. Instead, determinism posits that all events, human actions and decisions included, are ultimately set in stone by the laws of physics. It would therefore follow, as Laplace observed, that:

> An intelligence knowing all the forces acting in nature at a given instant, as well as the momentary positions of all things in the universe, would be able to comprehend in one single formula the motions of the largest bodies as well as the lightest atoms in the world.[22]

Laplacian determinism would eliminate human agency and free will and therefore stands opposed to a Catholic understanding of the cosmos and human beings. The good news, however, is that contemporary physics suggests Laplacian determinism simply is not true. The best current work in quantum physics suggests that at the physical level, events are nondeterministic, throwing a measure of indeterminacy into all events. The bad news is that the specter of determinism nevertheless continues to haunt many contemporary conversations about human agency. While Laplacian determinism might be false, the general idea that our lives are fixed by physical laws remains. For the determinist, our actions are not fundamentally caused by *us* but rather by forces beyond our control. Genuine human agency is a bad fit for the deterministic picture. If there is no genuine human agency or if human agency is but a pale shadow of what we typically imagine it to be, then once again, it becomes difficult to

[22] Pierre-Simon Laplace, *Essai Philosophique sur les Probabilités*, introduction to *Théorie Analytique des Probabilités* (Paris, 1820); trans. F. W. Truscott and F. L. Emory, *A Philosophical Essay on Probabilities* (Dover, 1951).

differentiate human agency from AI-powered modes of engagement with the world.

There is one last pair of theories to consider: *rational actor theory* and *utilitarianism*. Rational actor theory is a philosophical and economic theory that says people act in order to quantitatively maximize the fulfillment of their desires.[23] Utilitarianism is a moral theory that says people should act in ways that lead to the quantitatively greatest good (or "utility") for the greatest number of people. We will say more about these theories in the next section, but here we will focus on how both theories are aligned toward outcomes, optimization, and efficiency. The rational actor efficiently and rationally acts to fulfill desires. Likewise, the utilitarian endorses actions that optimize utility for the greatest number of people. Rational actor theory and utilitarianism thus elevate those aspects of human agency that are central in AI development and, in doing so, conspire to make us lose sight of crucial differences.

Materialism, behaviorism, functionalism, determinism, rational actor theory, utilitarianism: Each of these influential theories undermines a robust understanding of human nature and human agency. None of them explicitly conflate human agency with the operations of an AI. Yet when these theories are embraced (either explicitly or implicitly), it can become difficult to identify what makes human agency distinctive. Can an AI application achieve human agency? We have argued the answer is no. But given dominant ways of understanding the human person, it should not be surprising that many should continue to imagine it can.

Agency is Not Optimization

The choices that are presented to us in life are not a test of our ability to optimize a utility function. Long before such beliefs began to pervade human society, we instead focused on relationships, community, and meaning. It is only relatively recently that people envisioned their choices

[23] For theological analyses of these forms of practical reasoning, see Mary Hirschfeld, *Aquinas and the Market* (Harvard University Press, 2018); Scherz, *Tomorrow's Troubles*.

in terms of the quantitative maximization of goods; previously they survived and sometimes flourished without recent economic notions of what constitutes "the good life."

While utilitarianism brings free-market cost-benefit analysis to the realm of ethics, we should not be fooled. Just as market economics and capitalism attempt to maximize return on investment, utilitarianism attempts to maximize a return of pleasure. We pay a cost for pleasure in the form of labor, hoping to receive a greater amount back, typically in the form of money, which can then be turned into pleasure, whether through consumer goods, entertainment, power, etc. The abstract quality of money, which becomes a source for other goods, is worth acknowledging because it then becomes the locus for notions of optimization and maximization. We thus can be lured into focusing on maximizing money as the single good.

Yet there are many goods. Aquinas argues that for humans, "good" is too varied for the instinctive behavior found in animals.[24] We are not born with the instinctive knowledge to eat certain plants and not others, but rather learn what to eat, and not eat, from our cultures as we grow from children into adults. Furthermore, there are many other goods besides those required for survival; there are also the social goods of community and mental goods, such as ultimate truths about nature and God. This vast assortment of goods requires not a programmatic sense of "good" but rather a large abstract category of "good," within which specification is possible based on context. Utilitarian philosophers have agreed there are a plurality of goods, and in such a world, "optimization" for just one good can become not just counterproductive, but downright dangerous, especially if powerful AI systems are involved.[25]

---

[24] Thomas Aquinas, *Quaestiones disputate de veritate*, q. 25, a. 1, ad 4, trans. Robert W. Mulligan, James V. McGlynn, and Robert W. Schmidt (Henry Regnery Company, 1952–54), online edition ed. Joseph Kenny, isidore.co/aquinas/english/QDdeVer25.htm.

[25] See G. Arrhenius, "An Impossibility Theorem for Welfarist Axiologies," *Economics and Philosophy* 16 (2000): 247–266; Peter Eckersley, "Impossibility and Uncertainty Theorems in

Optimization is not an ideal approach to achieving a variety of goods, especially in unevenly situated contexts. The algorithmic approach to maximization often ignores appropriate context, and it can become impossible to balance between multiple goods and contexts. Moreover, it is generally difficult to translate a variety of goods into a single, commensurable scale. As humans are embodied images of a God who is both Love and *Logos*, both Unity and Trinity, the complexity of the human condition should not be overlooked. Complexity makes optimization difficult or impossible. However you look at human beings, we are not about only one trait. Any attempt to reduce a human being to something optimizable is an attempt to reduce our humanity, period.

While there are many goods in life, the greatest is God. In our relationship with God, no "maximization" or "optimization" is possible. We can grow toward God, but we cannot mathematically optimize our relationship with God. Nor can we do so in our relationships with other humans. This is a second mistake to recognize: Optimization is not applicable in all contexts. Relationships are not efficient, nor should they be. Good relationships require love, and love requires time. Optimizing time by minimizing the amount of it we share with others violates love because it subordinates this ultimate end of all human beings to some other measure, in this case time. Even if we "optimized for being as loving as possible" to make people feel better or maximize a person's prospects for entering Heaven, the instrumentalization would betray the love by making it a mere means to some other end. Love wills the best for others, and any attempt to maximize this, while perhaps seemingly beneficial, in fact detracts from both the means and the end.

A third error is thinking that "good" is a product we produce via our agency. Utilitarian movements such as "effective altruism" can easily fall into this trap by focusing on the production of "good" as though it were a material product from a factory, ignoring the roles of process, motivation,

---

AI Value Alignment (Or Why Your AGI Should Not Have a Utility Function)," *arXiv.org* (2019), arxiv.org/abs/1901.00064.

and authenticity.[26] For example, many successful small-scale poverty alleviation programs validated by randomized controlled trials fail when they are scaled up because their success depended not just on their methods but on the virtues, dedication, and creativity of their founders.[27] This is not to impugn the motivations of some who follow such ideas, but merely to point out an obvious pitfall. The point is not only to give aid, but to love the other in need—to grow and share in virtue not by focusing on oneself but by willing the good of the other. As images of God, we grow in God's likeness through the act of loving others. Growth in virtue means growth in our capacity to love, our capacity to will the good of the other.

This is the core of human agency: to will the good for everyone. As an orientation of our volition, it is not so much a capacity to optimize as a path to walk. It is a journey that can only move at the speed of life, as we live it in every moment. Running on this path will get you there no faster. When we choose love, we use our freedom and agency for the purpose for which it was made. We thereby realize our humanity by becoming more humane. Agency is the means to this end, the end of love itself, which is also our return to God, who is Love.

As John Paul II suggests, promoting a "culture of life" facilitates individual agency toward its loving end.[28] For Francis, this means stepping away from a "throwaway culture" that sees human beings as economic constructs—assets or liabilities to exploit or discard.[29] "Perfect love drives out fear" (1 John 4:18). Thus, love is not merely the goal of agency but also its enabler. As we ponder the role of AI in society, we would do well to

---

[26] For the ideas behind effective altruism, see Peter Singer, *The Life You Can Save: How to Do Your Part to End World Poverty* (Random House, 2010).

[27] Jeff Tollefson, "These Experiments Could Lift Millions Out of Dire Poverty," *Nature* 606, no. 7915 (June 22, 2022): 640–642, doi.org/10.1038/d41586-022-01679-y; Scherz, *The Ethics of Precision Medicine: The Problems of Prevention in Healthcare*, 131.

[28] John Paul II, *Evangelium Vitae*, §21, 28, 50, 77, 82, 86–87, 92, 95, 98, 100.

[29] Francis, *Laudato Si'*, §16, 20–22, 43, and *Fratelli Tutti*, October 3, 2020, §18–21, 188, vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20201003_enciclica-fratelli-tutti.html.

remember that our task is to make love ever more real in this world, to empower our agency and that of others. Insofar as AI might help us enact that noble end, we can be thankful for it. But we should not attempt to use any AI that directs us away from each other and toward things, or takes agency away from people, whether directly, indirectly, or through the magnification of fear. Every "no" protects a "yes," and the "yes" of human agency is God's love, the ultimate good that must be protected and sought after as life itself.

Part II
Specific Problems Related to AI and Agency

The first part of this book provided an overview of human agency and its differences from the way AI completes tasks. In this second part, we turn from these theoretical questions to ask how AI applications are concretely affecting the exercise of human agency today. The false understandings of human agency described in chapter 4 (behaviorism, functionalism, utilitarianism, and so forth) do not remain in the realm of theory or philosophical monographs. Instead, these models of the self are built into our technologies, embedding the technocratic paradigm and its false anthropologies into our everyday lives. These improperly designed technologies then go on to impair human agency. It is essential to understand the way AI affects our action. The coming chapters will explore some of the modalities of AI technology design that cause a negative impact on human agency in the realms of commerce, government, work, and the media. We will not cover every AI application or all its effects on human agency but seek instead to provide illustrations and points of departure for further discussion.

We should emphasize, however, that these chapters are not meant as a criticism of AI tout court. AI can be a tool for enacting the stewardship of creation that is an aspect of the *imago Dei*. Rather, these chapters are meant to be a critical analysis of some popular forms of technology in order to help inspire better designs. Later, Part III will explore what kinds of social formations lead to AI systems that support human agency.

CHAPTER 5

NUDGING, MANIPULATION, AND ADDICTION

AI is often used to manipulate human behavior, and frequently these manipulations draw on the insights of behavioral economics. This is a branch of economics developed to counter the notion that humans are perfectly economically rational actors who optimally weigh costs and benefits to maximize their subjective utility, which has long been a tenet of economic models, as chapter 4 described for rational choice theory. However, in the 1970s, economists and psychologists, most prominently Amos Tversky and Daniel Kahneman, used experiments to show that in reality, people's choices diverge from perfect economic rationality, often in predictable ways.[1] They called these predictable deviations "biases" and developed a long list of cognitive biases that has since been expanded by others. For example, in classical rational actor theory, people should be neutral between expected gains and losses, but in fact people are more sensitive to the costs of losing an object than to the benefits of gaining it—a bias called loss aversion. We want to stress that in behavioral economics, the term "bias" does not refer to injustice, as it does in other areas of AI ethics, just to predictable deviations from classical economic theory. Nor are these deviations necessarily irrational. Another school of theorists

[1] Daniel Kahneman and Amos Tversky, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica* 47, no. 2 (1979): 263–291, doi.org/10.2307/1914185; Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011). For a theological analysis, see Scherz, *Tomorrow's Troubles*, 128–131, 202–204.

prefer to think of biases as heuristics that promote effective action in the real world in which people lack perfect information, as opposed to the world of classical economic theory.[2] The important point for our purposes is the predictability of biases.

This predictability allows biases to be exploited by governments, corporations, and others to shape human behavior. The idea, popularized by Richard Thaler and Cass Sunstein, is that the architecture surrounding choices could be designed in such a way as to bias decisions in a certain direction, as will be discussed in more detail below.[3] These manipulations of behavior are called "nudges" because they are not constraints that overwhelm human freedom; people can still push back or choose differently. Nudges just tilt the likelihood of action in a preferred direction. They have become popular in policy tools, with the UK government even organizing a Behavioral Insights Team to design nudges for social ends.

Nudging is central to the current use of AI. Companies have used AI to design and implement nudges on users of apps and social media, building a new modality of business called "surveillance capitalism" that underlies social media and much of the Internet economy. An AI application's access to vast stores of data, inferential ability, and possibility of quickly testing different approaches allows for the design and personalization of nudging at scale. Moreover, insights from behavioral economics allows programmers to design applications that are dangerously addictive in order to increase user engagement with sites. These design paradigms aim to shape human action.

This chapter begins by investigating two questions. First, what does the predictability revealed by behavioral economics mean for human agency? Does it suggest a lack of freedom or a faulty rationality? Second, how do we ethically evaluate nudges? They are a growing aspect of contemporary

[2] Gerd Gigerenzer, *Rationality for Mortals* (Oxford University Press, 2008).

[3] Richard Thaler and Cass Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Penguin, 2009).

society, so it is important to inquire into their implications for responsible agency. The remainder of the chapter will then describe the effects of the exploitation of psychological theories in surveillance capitalism and related addictive technologies.

Predictability and Human Freedom

Nudging aims to influence our choices. Depending on the technology, on its efficiency and mode of operation, nudging may or may not undermine the freedom with which we choose. Chapter 2 articulated the nature of free will. But here new questions arise, for nudging avails itself of certain "mechanisms" to influence our will. The term "mechanism" implies that there is a way in which these influences regularly work, i.e., that there are rules of human behavior that such a mechanism can use. This implies that human behavior and human choosing are predictable. People who strategize political propaganda and advertise merchandise have long known how to employ such predictabilities, based on the regularity of our behavior at the level of the population. Current machine learning can fine-tune such predictability even at the level of the individual, identifying the patterns of our behavior in detail and consequently promising high success in manipulating it. The degree to which this is possible once more raises questions about the very reality of free will. How free are we, really?

Statistical predictability is not in and by itself an argument against or an obstacle to free will. There are two reasons for this. First, in the nineteenth century, the early analysts of social statistics discovered social regularities in seemingly free action like crime and suicide.[4] These regularities seemed to many people to suggest that crime is not freely chosen. Yet there is no reason to think that the individual acts described by these statistics are not free. Even Thomas Aquinas noted that, though we may be in a state of

[4] For this history, see Francois Ewald, *L'etat Providence* (Grasset, 1986); Theodore Porter, *The Rise of Statistical Thinking 1820–1900* (Princeton University Press, 1986); Ian Hacking, *The Taming of Chance* (Cambridge University Press, 1990). For a theological discussion, see Scherz, *Tomorrow's Troubles*, chapter 6.

grace, there is a strong likelihood of falling into venial sin; short of special grace, we will indeed eventually fall. Nonetheless, the statistical predictability in question applies only to all cases collectively, whereas distributively, in each individual case, the person remains free.[5] Jesuits in the sixteenth and seventeenth centuries would compare this to the repeated roll of dice. If we roll dice at a large enough scale the results are also statistically foreseeable, yet each roll is causally independent from the others and remains free and unpredictable.[6]

Second, not every regularity is incompatible with free will. Recall that every faculty is a faculty *for something*, a *telos* or purpose. Free will, too, has a purpose for which God made it. Free choice must not be understood as aiming at arbitrariness; flipping a coin is not the paradigm of freedom. Freedom is the ability to pursue the good for which we are made, and in the case of rational animals the ultimate good is a unique, complete self-giving relationship to God. Embedded within this overall good are the goods of human nature, which include incentives tied to the sensitive faculties, such as emotions, and biological needs, such as food. Thus, we will, for example, eat regularly and predictably, unless there are overriding reasons to the contrary (a greater good, or perhaps a lack of food). We will do so rationally, freely, *and* predictably. Moreover, in the regular pursuit of these goods we develop skills and habits, some of which become character traits. These also make us predictable, but this does not mean that we did not develop them freely or that we are not responsible for them. Nor are they contrary to our freedom. The skill with which I play tennis means that I do not consciously deliberate about every move that I make, but this does not mean that I play tennis involuntarily. Instead, habits and skills free me to deliberate and decide on more important things

[5] Thomas Aquinas, *De veritate*, q. 24, aa. 12 and 13. See also Peter Lombard, *Lib. Sent*. I, d. 25, c. 5; Augustine, *De civitate Dei* 22.30.

[6] Sven K. Knebel, *Wille, Würfel und Wahrscheinlichkeit: Das System der moralischen Notwendigkeit in der Jesuitenscholastik 1550–1700* (Felix Meiner, 2000); Tilman Ramelow, *Gott, Freiheit, Weltenwahl* (Brill, 1997), 129–139, 355–356.

and goals. They also remain part of the overarching good that is the fulfillment of freedom. Such skills include moral virtues; because of such virtues it is predictable that a saintly person will not do outrageously immoral acts. But this implies greater, rather than lesser, freedom on the part of the saint. Because freedom is a teleological notion, there is an asymmetry here: Someone steeped in vice is less free, while virtue sets us free to do the good for which we are made.

All of this said, some regularities may indicate a problem. Factors external and contrary to our pursuit of the good may derail us, preying upon our vices.[7] Multitasking, stress, and surprising situations may throw us back into bad habits. In those cases, only a wise person will take a step back and correct any bad inclinations in view of the larger good.[8] Most virtue theorists have assumed that the perfectly wise are in the minority, which makes us susceptible to manipulation. Any nudging that would reinforce such influences that alienate us from our true good would be against our freedom. Unfortunately, not only does current technology allow those with power to target the worst in us for the sake of manipulation, but it also has created a culture in which precisely these dangers to our deliberative choice are aggravated: multitasking, rapidification, stress, and surprising situations that ask for quick responses. All these circumstances threaten to make our life into a series of knee-jerk reactions that have nothing to do with the rational deliberation or the building of sustained virtue that free will presupposes as a teleological faculty.[9]

---

[7] In cases of *akrasia*, the passions may prevent us from the proper application of universal, rational principles, removing one of the necessary conditions for free acts. Thomas Aquinas, *De veritate*, q. 16, a. 3c.

[8] Aquinas, *ST* I-II, q. 9, a. 5; *Summa contra gentiles* III, c. 85, 21-22. *De veritate*, q. 22, a. 9, ad 2. Theologically speaking, this is one of the reasons why we need grace, not just for our will, but for other faculties as well.

[9] Further aspects and references in Anselm Ramelow, "AI Algorithms and Human Free Will: How New Is the Challenge?" in *The Promise and Peril of AI and IA: New Technology Meets Religion, Theology, and Ethics*, ed. Ted Peters (ATF Press, 2025), 385–398.

Ethical Distinctions

These considerations, and those from earlier chapters, suggest criteria for judging the morality of a nudge. There are three aspects to consider: the ends sought, the modality, and the relational context. Two caveats: First, here we are primarily considering nudges that explicitly consider the biases identified by behavioral economics to manipulate the choice architecture of a decision. There is a broader way of speaking of nudges that we are not using; this broad use of the term might suggest any effort to shape human action, including moral formation or explicit attempts at persuasion. Second, behavioral economics is not only used for purely economic analyses but has been deployed in a broad suite of policymaking, including for sustainability and public health goals. Thus, a discussion of behavioral economics leads beyond the sphere of the market. However, we focus here on the market, the context in which AI is most likely to be used for profiling people and targeting them for nudges, and for identifying likely situations for nudging, as will be discussed regarding surveillance capitalism.

The first ethical question concerns the ends sought by nudging. Thaler and Sunstein describe their project as a type of "libertarian paternalism"; they want to use nudges to promote what they see as people's good, but in a way that does not deny freedom. The key here is that their goal is to encourage human goods and flourishing. Thus, they suggest setting the default option for retirement savings contributions at the maximum amount as a nudge to encourage savings for retirement (taking advantage of a bias toward maintaining rather than changing the status quo). In this case, preparation for the future is a prudential good. Others have shown that they can nudge people toward voting by showing them that their social media contacts have voted (exploiting the bandwagon effect).[10] This

[10] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, et al., "A 61-Million-Person Experiment in Social Influence and Political Mobilization," *Nature* 489, no. 7415 (September 2012): 295–298, doi.org/10.1038/nature11421.

nudge encourages people to fulfill their democratic ethical obligation. In contrast, other nudges encourage vices like consumeristic shopping or developing an addiction to an app. Nudges directed toward human good are more likely to be ethically good, whereas those aiming at vices are certain to be ethically bad.

The second major category to consider is the modality of nudging. Nudges can occur in many ways, and their exact form and which biases they target matter. Here, we will give three examples of the many modalities of nudging. The first occurs in situations in which some kind of default option is inevitable, but any option chosen will shape the choice. For example, shoppers are more likely to select whatever foods are on the endcap of an aisle at a supermarket or at eye level in a cafeteria because of an availability bias; we are more likely to pick things that are immediately within our sight or memory. The problem—*something* has to be put there—is unavoidable. Whoever is organizing the store simply has to choose something to put in these positions. The question is: Should it be candy or fruit? A second example of this kind of nudge has already been mentioned: the default option for new employees' retirement savings contributions. Because of a status quo bias, people tend to just leave whatever the default option is in place. Therefore, it matters whether a new employee's default retirement savings is set at the maximum or at nothing. In such situations, a business has to make that decision.

A second modality concerns framing effects. The way someone communicates options can sway decisions. Thus, if a doctor describes a treatment in terms of negative outcomes (20 percent chance of failure), patients are less likely to choose it than if she describes it in terms of positive outcomes (80 percent chance of success). Patients make the differential choice even if the different framings describe the same result (an 80:20 likelihood). Framings can therefore be deliberately chosen to encourage a certain outcome.

A third nudge can occur through individualized prompts to action that can affect people at a subconscious level. For example, mapping programs

can direct a person on a route past a restaurant at lunchtime to nudge them to stop there. Or social media algorithms could preferentially introduce posts by our friends saying that they voted in order to encourage others to vote. These kinds of individually targeted actions are the most prominent way that AI is used in nudging. AI creates profiles to determine the most effective kind of nudges, and its algorithms can deploy them at times that their users are most vulnerable.

The last major consideration is the relational context of the nudge. Is the nudge occurring in a face-to-face interpersonal encounter, or is it introduced by a faceless bureaucratic organization? Is it, as in the case of the doctor-patient encounter, done within a trusted relationship of fiduciary authority, or by someone with responsibility for the individual nudged in some other way, like a parent for a child? Or is it carried out by an anonymous business attempting to maximize sales? This relational setting can determine how exploitative a nudge might be.

Ethical Analysis

Three ethical considerations help to determine whether a nudge respects human agency. Considering the first modality, in a situation in which there is an inevitable default, then there is no escaping the fact that someone, usually someone in a bureaucratic organization, is going to shape the outcome. It is better that it is done in a way that is more likely to encourage human flourishing or at least help to avoid negative outcomes. Thus, we should place fruit at eye level in the school cafeteria and make the default saving for retirement the maximum rate. People can still alter these choices if they feel strongly, and thus exercise agency, but the default encourages a prudential outcome. In contrast, the nudge would be immoral when it provides temptations to vice or pushes toward a negative outcome.

Framing effects require a more complicated analysis. They are clearly ethically bad when they deliberately manipulate a person's understanding of reality in the direction of falsehood. For example, misinformation on

social media can distort a person's understanding of a situation through framing effects or half-truths without necessarily telling outright lies. Or framing effects can encourage consumerism. In contrast, deliberate framing effects can be legitimate when done in interpersonal encounters with trusted authorities. Thus, Gerd Gigerenzer and others have argued that patients are already looking to the doctor for a recommendation and reasonably take the framing as communicating that recommendation.[11] Being influenced by the doctor's framing is therefore the result of trusting her, a trust that emerges from a deeper relationship shaped by the fiduciary nature of the doctor-patient encounter.

One could see the ethics of such framings as falling under the broader categories of the art of persuasion and rhetoric, which are necessary for human society, even though they can be misused. Yet, even regarding legitimate relationships of authority, it would be better for the person in authority to give a fuller explanation of the different options and their advantages and disadvantages, rather than merely communicating their recommendations through framing. For example, health educators design graphic materials on the risks and benefits of procedures so that patients can better understand outcomes.[12] Alternatively, the doctor could simply take the time to walk through different treatment options using a variety of framings. Such an explanation better supports a rational decision on the part of the patient. Since human freedom is based on rational agency, such steps encourage freedom. Moreover, helping people decide and exercise responsible agency supports the development of the virtue of prudence.

The last modality is the individualized prompt to action. Such prompts are almost always illegitimate, and would be even if they were being used for good ends by a legitimate authority. Even in the best cases, they attempt

---

[11] Gerd Gigerenzer, *Rationality for Mortals*.

[12] Gerd Gigerenzer, Wolfgang Gaissmaier, Elke Kurz-Milcke, Lisa M. Schwartz, and Steven Woloshin, "Helping Doctors and Patients Make Sense of Health Statistics," *Psychological Science in the Public Interest: A Journal of the American Psychological Society* 8, no. 2 (2007): 53–96, doi.org/10.1111/j.1539-6053.2008.00033.x.

to supplant the individual's responsible agency, subtly pushing them toward action rather than persuading them or allowing them to consider the different options available. More often than not, however, they are carried out by algorithms designed by large corporations to maximize profit. They are attempts to evade our deliberative capacity, which, as we have noted, is essential to our rational agency. This is an insult to the dignity of human agency. Designed to take advantage of an individual's specific weaknesses, they create very powerful stimuli that are extremely difficult to even notice, let alone overcome. In so doing, they treat the person as a behaviorist stimulus-response machine, partaking in an anthropological perspective inimical to the *imago Dei*. Such nudges are thus unwarranted.

Surveillance Capitalism

One of the most prominent uses of nudging occurs in the economic model of surveillance capitalism.[13] This strategy of economic accumulation combines nudging with the ubiquitous surveillance of our activities enabled by the vast amounts of personal data that everyone leaves in the digital environment, as well as the power of AI to swiftly analyze this data. Surveillance capitalism refers to a business model, popularized by search engines and social media platforms, in which the products and services are given away without a direct monetary cost. Search engines and social media platforms do not exchange their goods and services for money, as in a traditional commercial relationship. Rather, the primary exchange between these companies and their users is one of information for information. Individual users get Internet search results, turn-by-turn directions and real time traffic to a destination, or silly cat videos, and, in exchange, tech companies gather information, including search histories, geolocations, personal preferences, and tastes. Tech companies then use these stockpiles of data to identify patterns of personal behavior that allow

[13] This section draws on the analysis of Shoshana Zuboff, *The Age of Surveillance Capitalism*.

companies to predict future action. These "behavioral futures" are used by the analyst or sold to other companies to enable them to design nudges that will manipulate behavior. These nudges include the most precisely targeted advertisement opportunities in the history of the advertising industry. These precise advertising opportunities are then sold, in a traditional service-for-money exchange, to advertisers. Such a system betrays all the problems of nudges in relation to human agency discussed in the last section.

Further, individuals and collectives should not be reduced to objects or commodities whose actions or inactions are analyzed and influenced in quantitative terms—e.g., which consumer goods this person or social group can acquire, how much this individual or collective can be influenced to buy market products. Objectification and commodification of human beings betrays the materialism that characterizes the throwaway culture that Francis wrote "affects the excluded just as it quickly reduces things to rubbish . . . [and] affects the entire planet."[14] The search for what is humanly valuable and for what has meaning is challenged by social contexts and dynamics that lack meaning and aim only at pursuing and increasing profit to the benefit of very few people in positions of power.

The technologies built for surveillance capitalism can also be put to coercive uses by authoritarian regimes, which show their aggressive logics and practices. This also highlights the resilience and adaptability of this form of capitalism. Within diverse political and social arrangements, surveillance technology is equally used as a tool to establish and reinforce forms of social and economic control and to further enrich those who are already rich and powerful. Hence, the scholar Shoshana Zuboff refers to surveillance capitalism as the new frontier of power in global society.[15]

The focus on individual behavior reinforces the neoliberal understanding of agency that assumes that individuals, instead of the wider society, should be responsible for their own welfare. Hence, in case of any

[14] Francis, *Laudato Si'*, §22.
[15] Zuboff, *The Age of Surveillance Capitalism*.

hardship, both the responsibility and the blame fall on the individual. There is neither intentional awareness nor critical assessment of exclusionary dynamics and of social factors like racism, social inequities, marginalization, and discrimination. This privatization and individualization of the social realm is accompanied by an emphasis on the need to limit and curtail the role and responsibilities of civil society toward its citizens. By building the technological infrastructure offered by AI, these social forms enter into our tools, shaping our surroundings. The enhanced surveillance serves to disempower citizens and hinder individual and collective forms of resistance. Finally, these forms of control generate social distrust and lead moral agents to experience further marginalization and disengagement.

Addiction

One way to understand this business model is to argue that the public are not the customers of tech companies, but the product, sold by tech companies to advertisers. Such an argument raises the uneasy connotation of users being "sold" as objects and thus begins to highlight some of the concerns of human agency inherent in this business model. But the objectification of technology users actually hides some of the most perverse elements of this system.

By removing the material costs to users of these systems, tech companies have reduced the traditional barriers and made their products enticingly easy to use. These products are not without cost, but the costs are rendered invisible to the user, and thus the user does not hesitate to consume one more search query or continue doomscrolling on their social media feed.

This strategy would be harmful enough, but the business model for these platforms measures actual material profit in the number of advertisement spots that can be sold, which directly correlates to the amount of time a user spends on a given platform. Thus, tech companies are incentivized to maximize user engagement on the platform to drive up

profits. Put simply, the business model for most digital platforms creates perverse incentives to make their platforms as addictive as possible.[16]

In the early days of social media, tech companies hired experts in psychology to create user interfaces that kept individuals on the platform. Features such as infinite scrolling prevent users from reaching a natural stopping point where they can make the conscious decision to put down the device. Features such as swiping down to refresh delivers the same dopamine release as a slot machine in a casino. In other words, tech companies purposefully designed their platforms to be addictive in order to maximize revenue. Addiction, of course, is perhaps the most extreme form of the degradation of human agency.

This situation has been further exacerbated by algorithms tuned to maximize user engagement. AI systems can process and discover patterns in user data that the user may not even be aware of, and then capitalize on that data, not for the user's own good but for the maximization of engagement on the platform, and, in turn, as noted, the company's profit. As the power of AI tools increases, the ability for these algorithms to exploit human weaknesses increases. Today, some AI platforms are so capable of interacting and sounding like a human that the platforms themselves may seem to replace genuine human relationships. Recent studies have shown that individuals are spending less time together than ever before, and news stories indicate that some people are turning to AI bots instead of human therapists or romantic partners.[17]

Beyond the individual user, this perverse business model has also created a social expectation that digital products *should* be free. This makes it extremely difficult for companies that would prefer a different business model to be able to compete. It also socializes the dependence on addictive technologies. Social pressure can also be a powerful force for maintaining

[16] For this discussion, see Natasha Dow Schull, *Addiction by Design: Machine Gambling in Las Vegas* (Princeton University Press, 2012); Adam Alter, *Irresistible*.

[17] Derek Thompson, "The Anti-Social Century," *The Atlantic*, January 8, 2025, theatlantic.com/magazine/archive/2025/02/american-loneliness-personality-politics/681091.

an addiction. The network effects of an established platform, even one that its users know to be harmful, makes it impossible to abandon the platform without also abandoning the network that is constructed on top of that platform.

Ethical Analysis of Surveillance Capitalism and Addictive Technologies

Catholic appeals for an ethical response to surveillance capitalism must begin with a restoration of human dignity. The business model that turns users into products necessarily strips them of their fundamental dignity. However, the AI systems discussed in this section do not merely strip the inherent value of a human life and measure it in mere material terms, as so many other capitalistic systems do. By reducing relations between tech companies and users to information-for-information exchanges, technology companies and the algorithms they deploy reduce the person even further to bits of information. These mere bits of information are both an abstraction of the true human being and the very tool that exploits the human user. Restoration of human dignity involves a recentering of the customer as an active agent in the commercial exchange of goods.

Further, by stressing that all human beings, regardless of their nation of origin, race, gender, and social status, are entitled to specific rights and freedoms, human rights offer a universal framework, with possible legal implications, that aims at protecting vulnerable individuals and collectives and advancing justice in society. As Gostin and Meier stress, "Instrumental to human dignity, rights seek to address basic needs and frame necessary entitlements to uphold a universal moral vision, reflecting what a person is entitled to have, do or receive."[18] Hence, for them, "human rights establish a normative foundation for contemporary understandings of global

[18] Lawrence O. Gostin and Benjamin Mason Meier, "Introduction: Global Health and Human Rights," in *Foundations of Global Health and Human Rights*, ed. Lawrence O. Gostin and Benjamin Mason Meier (Oxford University Press, 2020), 6.

justice."[19] However, the difficulty of pursuing violations of human rights and implementing forms of national and global accountability based on international declarations further limits the social impact of advocating for and relying on human rights in confronting surveillance capitalism.

While human rights allow for a focus on human agency and point to discriminatory practices, negotiating human rights demands addressing the complex relationship between two factors. On the one hand, human rights as articulated in international documents argue for a new vision of politics based on transnational governance and for rethinking the social contract by stressing equity. On the other hand, the history of human rights highlights how the formulation and promulgation of these rights have been the expression of colonial powers. Decolonial and postcolonial critiques highlight the limits of this ethical resource, which is inseparable from the history of colonial oppression experienced by countries in the Global South.[20] Hence, these criticisms highlight limitations of human rights in protecting moral agents from the abuses that might characterize surveillance technologies. Human rights as a hermeneutical tool, however, could respond to these criticisms by expressing beliefs about human dignity within society with integrations and adaptations, while addressing the vulnerability of the poor.

Conclusion

While there are legitimate cases for certain types of nudges and while individuals can resist manipulative nudging, the overwhelming forces poured into manipulation are too great for most people much of the time. Humans have limited time, willpower, and attention, so the need to constantly resist these forms of manipulation is exhausting and even paralyzing. Most of these influences, further, are toward vicious outcomes like consumerism. They shape a society of ongoing temptation that

[19] Gostin and Meier, "Introduction," 6.

[20] Nelly Wamaitha, "The False Promises of Progress: Human Rights and the Legitimization of Inequality," *Journal of the Society of Christian Ethics* 41, no. 2 (2021): 297–314.

undermines responsible agency. If we are concerned about supporting agency, then we need systemic change, or at least regulatory action that prevents problematic nudging. This approach does not demand a complete rejection of nudges. Organizations with fiduciary relationships can use behavioral insights in situations where there will be an inevitable default. However, they should seek to engage a person's reason, to allow the person to participate in the good of the choice, whenever possible.

Chapter 6

Lost Opportunities for Agency Through Deskilling and Algorithmic Governance

While the AI applications discussed in the last chapter attempted to undermine agency through direct manipulation, AI can affect agency in less direct ways. In some cases, AI applications merely remove the opportunity for agency and the prerequisites for the individual to effectively exercise agency, as with deskilling and algorithmic governance. In deskilling, the person either never acquires or fails to maintain the habits and skills necessary to act well because many activities are taken over by machine. Under algorithmic governance, the individual citizen as well as the government worker lose the intersubjective social context within which they can participate in public affairs. In each case, the individual's agency is subtly undermined.

Deskilling

The history of humanity is in part the history of technological development, as humans constantly innovate to extend our abilities and solve our problems. By developing new tools, people have expanded the ability to enact agency and created new forms of responsibility. Yet, these new technologies produce changes in the societies and people who rely on them. One feature of almost any technological innovation is that a task

that used to be important or necessary is obviated, and since we are no longer forced to perform the task, the requisite skills tend to fade away. This phenomenon is called "deskilling."[1] Is deskilling another one of the negative consequences of technology? Or, by relieving us of the burden of learning and practicing some irrelevant skill, is this yet another way that technology makes our lives better? Like technology as a general phenomenon, deskilling is neither good nor bad in itself. We can find examples of when it is generally good, such as when the ability to forge metal tools replaced the skills of flint knapping, and we can find examples which many people find harmful, such as how widely available, cheap, processed food has led to the loss of skills in growing, preserving, and preparing nutritious food.[2] But when exactly is deskilling a good thing? And when is it a bad thing? And what marks the difference?

These turn out to be surprisingly complex questions. It is often difficult to tell if a technology has taken away some skill that is useful or necessary for us, and still harder to predict whether a technology will, in the long term, eventually take away something important. It often seems as though we are living through an endless trial-and-error process of discovering whether a particular technology has been harmful for us. Yet, perhaps at this historical moment, we who have experience of both a pre-digital and post-digital world are in a privileged position to judge the value of what is endangered by some aspects of AI-driven deskilling.

An instructive example of someone similarly caught at the institution of a new technology, that of writing, comes from Plato's dialogue the

[1] The term deskilling was first applied in labor contexts in reference to processes like the Taylorism discussed in chapter 3. Workers' tasks were reduced from complex craft manufacture to simple assembly line procedures. See discussion in Harry Braverman, *Labor and Monopoly Capital* (Monthly Review Press, 1998). Scholars of technology have expanded the term to encompass moral skills and virtues undermined by new technologies like AI, as in Shannon Vallor, "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character," *Philosophy & Technology* 28, no. 1 (2015): 107–124, doi.org/10.1007/s13347-014-0156-9. See also Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §67.

[2] Michael Pollan, *Cooked: A Natural History of Transformation* (Penguin Press, 2013).

Phaedrus.[3] As always, the dialogic nature of the work makes it impossible to discern to what extent Plato is committed to the views expressed, but his character Socrates makes the case that the written word is not an ideal way to communicate truth. The conclusion seems to be that the advent of the technology of the written word is an example of harmful deskilling, because the ease of writing down our thoughts makes it less likely that we will be willing to endure certain mental processes that, according to Socrates, are arduous but at the same time necessary for the best kind of life. It is tempting to ridicule Plato on this point, as it is very difficult to imagine a modern-day good-faith argument that humanity is worse off because we can write things down.

But we think we should draw the opposite conclusion here. Given the many cultures that thrived primarily through oral traditions, including at least some Plato would have known about or encountered,[4] it is only in hindsight that we can think more clearly about the benefits of the written word and the limits of oral tradition. This example, therefore, only emphasizes how difficult it is to predict whether deskilling will turn out to be positive or negative.

What Does Deskilling Mean for Us?

"Deskilling" is the most common way to refer to the phenomenon in question, but it can be misleading because it implies that technical skills are the real issue. It is perhaps true that the loss of technical skills provides the most obvious examples, but this has the effect of making it seem that deskilling affects only the part of the human mind responsible for developing practical skills. But as the very name "artificial intelligence" suggests, we are now dealing with a fundamentally different kind of

[3] Plato, *Phaedrus*, 276, 274c–277a.

[4] Walter J. Ong, *Orality and Literacy: The Technologizing of the Word* (Routledge, 1982). For example, today we marvel at the memory of those who lived prior to easily available mass-produced books and what that meant for their scholarship, as described in Frances Yates, *The Art of Memory* (University of Chicago Press, 1966); Mary Carruthers, *The Book of Memory* (Cambridge University Press, 1990).

technology, which threatens to displace different kinds of tasks. This makes the conversation about deskilling much more difficult, because now the "skills" we lose may be higher-order cognitive skills, or cognitive skills that have moral salience, or moral "skills" (which may be properly called "virtues"), or a combination of those.[5] When deskilling touches on one of these categories, there are concerns that we may lose some capacity that is essential for a good human life. This is a clear place where the Catholic Church and "techno-optimism" clash, as any technology that diminishes our capacities for living a good life, as it has been defined by the Catholic tradition, is fundamentally bad.

We can begin to see the difficulty of making these distinctions when a task involves different categories of skills. Piloting aircraft shows how one profession has worked to protect some of its skills while still using automation for part of the job. As aircraft control technology has improved, the entire process of flying an aircraft, from taxi and takeoff to landing and taxiing again, can be automated. But despite this possibility, the autopilot of an aircraft is typically used only for the easiest and more boring parts of flying: the cruise phase. Precisely because takeoff and landing are difficult, they are not automated, helping pilots to retain their skill, which is thus available to be employed in unusual or emergency circumstances.[6]

Changing technologies cause changes in skill; this is no surprise, nor, necessarily, morally weighted. But virtues are excellences of human capacities, just as vices are perversions of them, and skills are likewise excellences, although not necessarily in a morally freighted way.[7]

[5] Vallor, "Moral Deskilling and Upskilling in a New Machine Age."

[6] Even in these easy and boring situations, automation can lead to disaster, with several crashes caused by pilots being unprepared and disoriented when forced to take control in emergency situations, e.g., Air France Flight 447 in 2009.

[7] For discussions of the relationship between virtues and skills, see Julia Annas, "Virtue as a Skill," *International Journal of Philosophical Studies* (1995): doi.org/10.1080/09672559508570812; Tom Angier, *Techne in Aristotle's Ethics: Crafting the Moral Life* (Continuum, 2012); Matt

Returning to piloting, an excellent pilot can fly an airplane in all sorts of conditions; this is a technical skill with moral salience, but not a moral skill or virtue directly. Piloting an aircraft is morally salient insofar as it involves morally important responsibilities for passengers' lives, encourages the development of bad character traits (e.g., by normalizing risk-taking behavior), or enables the pilot to develop virtues.

But some skills are more directly related to virtue and vice than others. Social skills are a prime example. While some poor social skills, such as thoughtlessness or rudeness, are clearly bad, the extent of their badness depends a great deal on context, as well as the expected developmental level of the individual in question. While rudeness might indicate malice, it also might merely indicate poor formation, and in that sense, it is more of an unformed capacity than a deformed capacity. This is often the difference between a rude child who has not yet fully developed self-control and does not have a complete grasp of social context, and an intentionally rude adult.

Returning to technical skills, these skills also reflect the technologies of the societies that they appear in. An economically and technologically advanced society has a much more extreme division of labor than a hunter-gatherer society or even a medieval society, for example. Division of labor makes it so that people specialize in certain skills at the expense of other skills. One might be a butcher, baker, and candlestick maker all at once, but it is more efficient to specialize. In that case, skills that one person might have combined long ago become disaggregated into different professions for the sake of economic gain. Thus, people lose the agentive capacity that would allow them to perform many tasks themselves but gain the capacity to easily access the goods produced by others.

Not only economic specialization but also technological change causes deskilling. For example, a "computer" used to be a person who calculated difficult mathematical problems. But the invention of the electronic

---

Stichter, *The Skillfulness of Virtue: Improving Our Moral and Epistemic Lives* (Cambridge University Press, 2018).

computer, then the handheld calculator, and now smartphones with a calculator app made this skill no longer necessary. And yet, the ability to perform mental math seems to be beneficial to the mind,[8] and thus removing this need to learn to do mental math, whether as a professional calculator or merely as a citizen doing mundane daily tasks, may remove a skill that is overall beneficial to individuals.[9]

More theologically, we might consider that the Bible implicitly acknowledges the reality that we cannot possess all skills through its consideration of vocations: Adam and Eve are to keep and till the Garden; Cain farms and Abel herds; successive generations in the early chapters of Genesis divide labor more and more finely with time. By the time of Jesus, the profession of carpenter was old, and we might wonder what it means that Jesus *was* a carpenter and not some other profession. He could have been many different things, after all—a potter, a farmer, a merchant, an aristocrat, a scribe, and so on. We know very little about Jesus's technical skills as a carpenter, nor what other skills he might have known incidental to his profession. Certainly, he was a skilled speaker and knowledgeable in the Torah. Morally, he was perfectly courageous, temperate, just, and prudent. But more than that, we might think of what technical skills Jesus did *not* develop: most of them.

This is further reason to believe that not all deskilling is intrinsically bad. It is just the way a highly complex society operates. As Paul notes in his letters (1 Corinthians 12:1–31, Romans 12:3–8, Ephesians 4:7–16, etc.), there are many and diverse gifts from God and therefore many and diverse roles to play in the Church, and yet we are all bound together in one body in Christ. Each of us will direct our action toward certain tasks as we

[8] Shinya Uchida and Ryuta Kawashima, "Reading and Solving Arithmetic Problems Improves Cognitive Functions of Normal Aged People: A Randomized Controlled Study," *Age* 30, no. 1 (March 2008): 21–29, doi.org/10.1007/s11357-007-9044-x.

[9] And here we might think of those who compare AI to a calculator and so realize that saying "AI is just like a calculator," even if it were a good comparison (which it is not), is still not necessarily a good thing.

discern our vocations, which will require forgoing the development of other skills due to our finitude.

When Should We Be Cautious About Deskilling?

But distinguishing between specific acceptable and unhealthy kinds of deskilling is very difficult, and, in fact, we are skeptical that there is any universal rule or principle that can articulate this distinction in the abstract. But we can develop practical guidance about it by framing the distinctions in terms of the level of caution needed, such that tasks might be grouped into 1) those we can have a low level of caution about automating or outsourcing to AI; 2) those we should have high caution about outsourcing; and 3) those we should never completely outsource, even when the skill involved in completing the task changes because of our collaboration with an artificially intelligent system. The question for tasks in the third category is only *how* humans should be "in the loop," not *if* they should be.

The first kind of tasks are those that are simple, repetitive, not requiring an extensive skillset, and only distantly related to the greater end of the activity to which they are contributing. For example, there is more art and agency involved in the initial design and physical testing of the plans for an automobile than in the repetitive action of tightening a bolt on an assembly line. This is an activity that is already situated within a technological context: The assembly line would not be possible without the mass production capabilities of technology. The repetitive action of the worker is confined to one detached role in the production process and is frequently alienated from the purpose of the practice.[10] A sign that a task falls into this low-concern category of deskilling is that there is not much room for growth in further appreciating the process to which the act is contributing: Though assembly-line workers are participating in a collaborative act of building a car, the extent to which they are engaged is

[10] For a discussion of the dissatisfaction of workers with this kind of overly repetitive assembly-line work, see Harry Braverman, *Labor and Monopoly Capital*, 21–24.

relatively minuscule, and they are detached from further involvement in downstream processes. In some cases, the person who is accomplishing these tasks is already brought into service for someone else's project; thus, the task has already been offloaded for a pseudo-automated process that only further separates the actors from the good achieved or product created. These types of skills tend to be rule-based, context-independent, and follow the same procedure each time. They are fully specifiable, do not engage the prudent judgment of the worker, and do not readily permit the engagement of the personality of the actor over that of the designers whose plans the worker is carrying out. These are skills that perhaps may be automated without a major loss. Nevertheless, we should scrutinize any automated processes and ensure that people who formerly did that task can find other forms of employment.

Further, there are many types of mundane tasks that can exhibit human care, achievement, and sacrifice. Something would be lost in the lack of appreciation for the physical therapy patient who successfully relearns the ability to tighten a bolt and can proudly return to his workshop, even if less precise with his tools than before. Something would be taken for granted if the friendly greeting of a grocery clerk who is otherwise a stranger is replaced by the automated robot who has no other option than to portray a simulation of patience for you.

The second kind of tasks are technical tasks that are closer to the purpose of an activity, that have a large scope for improvement, and that involve higher-order levels of thinking that could offer valuable challenges for human development. If there is a possibility of automating these tasks or outsourcing them to AI, we should have a high level of caution about how and when we do so. The calculator is a good illustration here. It is undeniably wonderful that calculators exist, but if we pay attention to the research, we understand that we should not simply get rid of mental math entirely. We must find ways to incorporate mental math into our lives even though calculators are always available.

The third kind of tasks are those that offer great scope to judgment, creativity, empathy, practical wisdom, and critical thinking. Most crafts, repair work, and professions fall under this heading.[11] This is the point at which "deskilling" becomes "de-virtuing," and we insist that these kinds of tasks should never be completely outsourced to AI. There may often be relevant AI tools, but we must focus on how those tools can be used to assist and enhance the essential human skills and virtues in these tasks and not replace them. A good example here are Large Language Models (LLMs). LLMs can be incredibly harmful from a deskilling perspective if they "save" us from the tasks of critical thinking, structuring ideas, expressing disagreements in a respectful way, etc. These are tasks that are important for the development of virtue, and outsourcing them entirely to LLMs is merely giving in to temptation. But it is not as simple as destroying LLMs, as many of us realize the amazing benefits of them, like organizing large amounts of unstructured data. Further skills are still needed even with the implementation of LLMs for these seemingly beneficial purposes. For example, human oversight and review are needed to root out biases and hallucinations (the so-far ineliminable problem of LLMs making up "facts" and references). If we use LLMs as a tool, this still involves the relevant capacities: We must still know what questions to ask, how to ask those questions, how to ask follow-up questions, how to evaluate the output, and how to use the output. But if we use them to avoid some task that requires virtue, like the proverbial student who asks an LLM to write an essay, we often will lose out on essential aspect of human development.

The crucial question for assessing dangers of deskilling involves how new technologies can turn us into either better or worse human beings who then go on to create a better or worse world. From this perspective, even the simplest, most repetitive task can be an opportunity to grow in virtue if it is done to serve God and others. Maintaining this intention deepens the moral significance of a mundane but good act, even the most

[11] Matthew B. Crawford, *Shop Class as Soulcraft: An Inquiry into the Value of Work* (Penguin, 2009).

minute good act. Even the smallest, most insignificant movements (opening the door for a stranger or hammering a nail) are ones that were performed by great saints with deep love to seek holiness, and even performed by Christ Himself. As St. Teresa of Calcutta was known to have said, "Not all of us can do great things, but we can do small things with great love." The choice that St. Teresa described is one that involves an interpretation and offering of the act that only a person can freely perform, not an AI. Yet, when accomplished with the cooperation of the person with grace, this intention behind a small good act can help to sanctify the world. We should not forget the moral value of even the smallest acts that as children we must master to understand the physical laws of the world and the coordination of our bodies; that at other stages of our lives, may bring to light our finitude and need for dependence; and that even in their most tedious, monotonous mode can be an invitation to the higher moral skills of patience, fortitude, and devotion to goods beyond the self.

Division of Labor, Complexification, and Relative vs. Absolute Deskilling

When it comes to deskilling there is one other matter worth considering. The increasing complexity of society causes a relative, even if not an absolute, deskilling. Absolute deskilling is deskilling measured against some objective, unchanging standard. This might be something measurable in terms of technical skills, such as computer programming or safe driving, but could be harder to measure when it comes to moral skills.

Relative deskilling, however, can occur even if an individual or all of society is staying at the same level of skill, or even increasing in skill, but not as fast as complexity (or demand for skill) is increasing. As skills ebb and flow among individuals and in society, division of labor, new technologies, and new complexities continue to develop as well. Can human skills keep up?

This increasing complexity was noted even in 1958, when Leonard E. Read wrote "I, Pencil."[12] In this brief essay, told from the perspective of a pencil, the narrator explains how incredibly complex it is to make a pencil, and that all the knowledge required does not and probably cannot reside in one individual. Instead, it is a massive social collaboration made possible by division of labor and markets, which together coordinate the production of something so mundane.

While making a small product like a pencil was incredibly complex in 1958, the world of today is vastly more intricate and impossible to conceptualize. It is so difficult, in fact, that economic and political shockwaves from the COVID-19 pandemic have reverberated for years, and continue to reverberate through the individuals affected by the pandemic, as their memories and skills have been changed through educational, social, and other impacts. Supply chains came crashing down as panic-buying, closures of production, and cessation of transportation all contributed to an unpredictable mess.

But as the world complexifies, technologies might also allow us to better manage this complexity. This is exactly where AI can contribute to humankind: helping to solve problems that are otherwise too large for humans to solve. In this way, AI is arriving at the right time to meet an emerging need. For example, every day there are satellites scanning the entire land surface of the earth and sending that data back to data centers for analysis of the effects of climate change and land use. No human could possibly look at all this data; in fact, not even a large group of people could look at it all. What AI can do is to look at it for us and then note differences that might be worth human follow-up. In this way, AI solves a problem that otherwise cannot be solved: detailed Earth-sensing analysis that upskills its human creators and users rather than deskilling them, and upskills them in a way that might be enough to keep up with the complexification of the overall growth of humanity, and its impact upon

[12] Leonard E. Read, "I, Pencil," in *The Cambridge Handbook of Classical Liberal Thought*, ed. M. Todd Henderson (Cambridge University Press, 2018), doi.org/10.1017/9781108242226.008.

the environment, across our planet. This is an empowering rather than deskilling use of AI. As our world grows more difficult to understand and manage, relative deskilling is unavoidable, and therefore AI does have a role to play in helping humanity meet our obligations to be good stewards of God's creation. At the same time, depending on AI to manage complex systems can create extreme fragility when it encounters radically new situations, so we should consider embracing forms of social organization that contain resilience and redundancy for critical functions.

In the context of human agency and freedom, we need to possess the skills that will make us truly free and our communities truly happy. These skills are context-dependent and can be moral and/or technical. As technology changes, techniques must change along with them, and as we become relatively deskilled, AI may contribute to helping us upskill ourselves and thereby be better prepared to tackle the problems of the future.

Algorithmic Governance

Governments and other large organizations are increasingly relying on AI and other algorithms to help carry out functions in areas such as national security and criminal justice. As mentioned in the last chapter, AI is contributing to oppressive authoritarian regimes through surveillance infrastructures that vastly increase government knowledge of the individual. This information then allows for programs like a social credit system that might result in detention or at least further limitation of the person's ability to travel and act. This is a problematic expansion of the coercive powers of the state. In many ways, these issues are not new, but AI is merely serving the expansion of the long-lasting secret police regimes of totalitarian states.[13] These coercive police systems in turn reflect an intensification of the general move toward surveillance as a mode of social

[13] Hannah Arendt, *Origins of Totalitarianism* (Harcourt Brace Jovanovich, 1973).

control in all modern societies.[14] They directly impinge on human agency through violence, intimidation, and the molding of character.

Yet not all government uses of AI are so coercive. AI is also used for activities such as the distribution of social welfare benefits like housing and food and for the detection of fraud. Many governments envision AI as a tool for greater efficiency and objectivity. Despite these promises, AI in governance raises deep questions for responsibility and democratic agency.[15] While less directly concerning than authoritarian uses, their effects are also more subtle, demanding a greater analysis. They tend to remove decisions from individual decision-makers and local communities. They also decrease the ability of citizens to directly engage with those who make decisions about their lives. Moreover, in many of these cases, AI directly impacts society's care for the most vulnerable, thus implicating Catholic social teaching's (CST) preferential option for the poor. All these features of algorithmic governance have implications for human agency.

The History of Algorithmic Governance

The desire for a more automated form of governance and decision-making is not new, arising out of a historical commitment to objective, quantitative methods like cost-benefit analysis. While they have roots in the bureaucratic methods discussed in chapter 3, these quantitative commitments have a distinctively American origin.[16] From at least the Jacksonian age, Americans have been fiercely committed to a democratic egalitarianism. In part, this commitment meshes well with the ideal of subsidiarity in CST and the US genius for independent organizations

[14] Michel Foucault, *Discipline and Punish: The Birth of the Prison* (Pantheon, 1977); Michel Foucault, *Security, Territory, Population: Lectures at the Collège de France, 1977–78*, ed. Michel Senellart, trans. Graham Burchell (Palgrave Macmillan, 2007).

[15] This section draws in part on insights from AI Research Group, *Encountering Artificial Intelligence*, 187–196, 235–241.

[16] This history is drawn from Porter, *Trust in Numbers*.

described by Alexis de Tocqueville (1805–59).[17] Yet these tendencies also lead to a deep distrust of expert decision-making in government. There is sometimes a desire for government decisions to not be made by unelected elites in the bureaucracy because such policy decisions seem arbitrary and unaccountable.

Still, not all policy decisions can be the subject of legislative action. People desired a way to make such decisions in a rule-governed and transparent manner. By the late nineteenth century, quantitative methodologies like cost-benefit analysis had arisen to meet the demand. Because they were quantitative, these procedures seemed objective, and because they could be clearly described, they seemed transparent, though as we will show below, these appearances were deceiving. Thus, government units like the Army Corps of Engineers began to use cost-benefit analysis to justify decisions about where to construct dams and levees. These tools spread over the course of the twentieth century, especially with the growth of environmental and health regulations, expanding into private and international organizations. Finally, these systems started governing judicial processes through techniques like strict sentencing guidelines and mandatory minimum sentences. This trend reaches its apotheosis in AI decision-making tools. They seem fully data-based with no dependence on arbitrary human whims.

Such tools are a rejection of dependence upon the prudence of officials. Decisions are not made by the mature judgment of individuals but instead by standardized, automated, and predictable algorithms. This tendency is clear in the fight over sentencing guidelines in the United States. Advocates became incensed over differential sentencing for the same crime by different judges, suggesting that the differences resulted from the arbitrariness of individual judgment.[18] In fact, most of these differences

[17] Alexis de Tocqueville, *Democracy in America*, trans. Harvey Mansfield and Debra Winthrop (University of Chicago Press, 2000), 489–492.

[18] Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein, *Noise: A Flaw in Human Judgment* (Little, Brown Spark, 2021), 13–23.

were due to substantive disagreements over the purpose of judicial punishment and specific differences in the crimes adjudicated, but the appearance of difference was enough to suggest the need for strict control of judges' responsibility.[19]

At first, other nations did not seek control over their bureaucrats in the same way. The United Kingdom, for example, with its history of an elite, public school–trained civil service, placed much more trust in officials. Even in these countries, though, trust in bureaucrats broke down with the spread of ideas like public choice theory and the broader rise of populism. For populists, government agencies, and corporate managers, quantitative methods seemed to free decisions from the ideological hold of bureaucrats. Thus, more and more decisions have been turned over to algorithms. These shifts have not had the intended effects of empowering democratic egalitarianism, though. Ironically, they have reinforced bureaucracy, decreased objectivity, evacuated accountability, reduced the agency of officials, and undermined democracy.

Reinforced Bureaucracy

As discussed in chapter 3, one of the great modern fears is the power of bureaucracy. Franz Kafka's works described, in admittedly surreal detail, the feelings of the individual agent who is seemingly powerless against the opaque authority of bureaucracy. The turn to algorithmic governance is in part a response to this fear, an attempt to provide a clear process through the system of unaccountable authority figures. However, that reading mistakes the dangers of bureaucracy by focusing on the individuals rather than the effects of the bureaucratic institution itself. Instead of decreasing the power of bureaucracy, AI in governance reinforces the spread of bureaucratic institutions.

The first reason for this is the mere fact that AI makes bureaucracy easier. One of the challenges of bureaucratic institutions is the difficulty of

[19] Kate Stith and José A. Cabranes, *Fear of Judging: Sentencing Guidelines in the Federal Courts* (University of Chicago Press, 1998).

keeping track and making use of the data its procedures generate. AI might solve these problems, making it easier to surveil every area of life. More and more information can be scraped from databases. Generative AI also makes it easier to deploy this data, and it can summarize vast quantities of material and make predictions. The mere existence of and access to such data is an opportunity for bureaucratic growth.

The increasingly AI-based bureaucracy will not necessarily be more transparent. Even in simple cost-benefit analysis there are difficulties in setting a quantitative cost for qualitative values like human life, environmental beauty, or ecological diversity.[20] While there are procedures for determining such values, ultimately their conclusions come down to human choices. As the next section will discuss, such opacity becomes greater with AI, whose conclusions are often unexplainable. This opacity might be acceptable if we were certain of its objectivity, but the results are often biased.

Finally, AI increases the problem of accountability. Placing responsibility for algorithmic mistakes is one of the greatest problems in AI ethics. Does responsibility lie with the programmer, the agency, the corporation, the user? As a later section will discuss, AI even enables the abandonment of responsibility by different agents. Thus, the opacity of bureaucracy becomes even greater because the citizen does not know who is responsible for her problem.

Far from undermining the arbitrary power of bureaucracy, AI will expand it. Bureaucratic procedures become easier, and thus they will tend to expand under algorithmic governance. Yet the bureaucracy will remain opaque and unaccountable. AI can lead to an even more Kafkaesque bureaucracy.

[20] Kristin Shrader-Frechette, *Science Policy, Ethics, and Economic Methodology: Some Problems of Technology Assessment and Environmental-Impact Analysis* (Springer, 2013).

Algorithms Are Not Objective: Distorting Human Agency Through Bias

Many early champions of AI argued that its objectivity could eliminate the biases against marginalized groups that affect institutions and their officials, as well as remove opportunities for corruption. While it may be tempting to think of algorithms as being clean, clear, objective, and unbiased, algorithms do not in and of themselves remove bias but rather can hide bias under a veneer of seemingly impartial math. This hidden bias is sometimes called "bias laundering," because human bias is hidden within a complex system, and by doing so looks "clean" of bias when it is in fact not.

On the other hand, AI can also detect and attempt to mitigate bias. Large data sets and algorithms can sometimes exceed human comprehension, and AI tools can be used to find biases in these data sets and algorithms. But, like all things involving AI—or technology more broadly—the detection and mitigation of bias can only occur if people actually do this work. Often systems are set up and then run with little supervision or human correction. There are numerous examples of automated systems running amok, where human intervention was sorely necessary but often came too late, if at all.[21]

When it comes to governance of complex systems, this tendency for bias to be hidden and/or ignored permits the agency of some persons to limit, undermine, or completely remove the agency of others. Poor governance with AI can distort the agency of an institution's subjects and

[21] See, for example, Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016, propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," Proceedings of Machine Learning Research Conference on Fairness, Accountability, and Transparency 81:1–15, 2018, proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf; and books such as Cathy O'Neil, *Weapons of Math Destruction* (Crown, 2017); Safiya Umoja Noble, *Algorithms of Oppression* (New York University Press, 2018); Ruha Benjamin, *Race after Technology* (Polity, 2019); Virginia Eubanks, *Automating Inequality* (Picador, 2019).

rob people of their freedom of choice, with AI as the instrument of that distortion and robbery. This can sometimes be intentional, but it can also be unintentional. In either case, the bias is unfair and ultimately harmful not only to individuals but also to society, and to those making and using AI systems, who will become targets of suspicion if problems like this proliferate.

Defusing of Social Conflict by Negating Human Agency

While, as previous sections showed, AI governance does not undermine the arbitrary power of institutions, the very appearance of objectivity can undermine any potential challenges to these institutions that might arise. Algorithmic systems are notoriously difficult to understand. With respect to human agency, a side effect of this difficulty is that when extremely complex AI systems make transparency difficult, then institutional officials can easily blame "the algorithm" whenever results occur that others might find distasteful. The AI system becomes a ready excuse. Lack of explainability can lead to either feigned or actual helplessness as an excuse to defuse interpersonal conflict over algorithmic decisions. Whether the person in the position of defending the algorithm is sympathetic to the victim or not does not matter: The algorithm can either be intentionally or unintentionally opaque, thus ending inquiries and appeals.

This makes the algorithm, or rather those who control the algorithm, the engineers or the top-level users, the de facto force in control. Everyone else becomes their willing or unwilling subjects. This trades the little day-to-day inconveniences of dealing with people made unhappy by a complex system for the much larger and much more intractable problem of a giant system that is largely inscrutable and controlled by unaccountable overlords. The effects on agency are clear: The people subjected to the system, whether external to the organization using the system, or internal to the organization in the form of employees "just trying to do their jobs," are disempowered and find themselves with less agency than they might

have had were the algorithmic system not there. Opportunities for a sympathetic ear to put in a good word are gone; human judgment becomes a flaw that has been removed. The desired mechanization of a human organization is achieved at last, with humans reduced to parts in an AI-powered bureaucratic machine. If "organizations are slow AIs," as some have said, then with the implementation of AI, the organizational impetus gains speed. The humanity found in slowness, deliberation, and a caring face is gradually removed, as the next chapter discusses.

Deskilling in Government

As discussed in the previous section, deskilling is a serious concern when it comes to the operation of AI in our world. While lack of appeal and agency harms those subjected to it, it also harms employees in the organization by deskilling them at their jobs, creating a growing dependency on the AI system, and ultimately possibly automating their jobs away. Suffice it to say again that in this context, if the desire for organizations is to shift from being "slow AIs" to being "fast AIs," then the removal of humanity and reduction of people to mere parts dehumanizes not only those subjected to the system, but those who are parts of the system itself.[22]

Ultimately, even those in charge of the system become trapped, not only as they become subject to their own externalized decision-making mechanisms and their effects, but also as they become reliant on the system for their own choices and agency. In this way it is reminiscent of the Kojèveian understanding of Hegel's master-slave dialectic, in which the master thinks they are in charge, but in fact they are completely dependent upon the slave, thus making the slave, in fact, "in charge" despite being ostensibly without authority.[23] Deskilling removes agency from

[22] For this point and the next section, see AI Research Group, *Encountering Artificial Intelligence*, 197–199.

[23] Alexander Kojève, *Introduction to the Reading of Hegel*, assembled by Raymond Queneau, trans. James H. Nichols, Jr., ed. Allan Bloom (Basic Books, 1969), 6–8; G. W. F. Hegel, *Phenomenology of the Spirit*, trans. A. V. Miller, analysis by J. N. Findlay (Clarendon Press, 1977), 112–118.

individuals and mechanizes it into a system, whether social or automated, alienating people not only from their labor but from each other, and, ultimately, through loss of agency, from themselves.

Undermining Democracy

The most ironic effect of the move toward quantitative governance is that in trying to decrease the power of elites in the name of democracy it may undermine democracy. It does so by decreasing possibilities for subsidiarity and thus participation. As we will discuss in more detail in chapter 11, subsidiarity is one of the central pillars of CST.[24] It is the core principle for how different social units like the family, government, and civil societal institutions relate to one another. This right relationship of such institutions is the original meaning of social justice.[25] The thrust of subsidiarity is that each social unit should be empowered to enact its proper function for the common good. The function of the family and local school board should not be usurped by a national government. Neither should the national government be deprived of its power to support the common good by coordinating programs for social welfare or the national defense that cannot be carried out by any other group. Subsidiarity argues that every action for the common good should be carried out at the proper level of society, with more centralized powers supporting and coordinating the actions of local groups.

The primary role of subsidiarity in CST has been to protect against improper centralization of power in order to support individuals' responsible agency in service to the common good. It is easier for more people to participate in serving the common good if more functions are carried out at the local level. For example, more people can participate in educational policy at the level of the school board than at the level of state

[24] Pius XI, *Quadragesimo Anno*, §79–80.

[25] Russell Hittinger, "The Coherence of the Four Basic Principles of Catholic Social Doctrine: An Interpretation," *Nova et Vetera* 7, no. 4 (2009): 791–838.

government, and NGOs allow more people to respond to social problems. Subsidiarity widens possibilities for participation.[26]

Subsidiarity therefore allows people to carry out their proper responsibilities. The family has a responsibility to raise children that is prior to the state. We have a responsibility to respond to the suffering people around us through actions stemming from personal encounters. We have responsibilities to act for the common good of our local communities. Subsidiarity is essential for allowing this participation and thus the fulfillment of our responsibilities.

By allowing more voices to participate at the local level, subsidiarity also enables democracy. Democracy places the weight of the common good on the people as a whole, a responsibility enacted not only at the ballot box once a year, but through everyday commitment to the common good. A network of more local organizations allows for this responsibility to be enacted.

The easier it is to centralize functions, the more pressure will be placed on subsidiarity, and AI makes it more likely that functions will be centralized. AI itself just works better on large populations, because it depends on statistical correlations. There will be more data to draw from, so its predictions will be more accurate. There is less likelihood of the outlier effects that occur in small populations. Additionally, it has so far required great resources and expense to train AI, which skews its development toward higher levels of government with more resources and more potential to realize an adequate return on investment. Unfortunately, such centralization will make such algorithmic governance less responsive to local variability—another, more pragmatic, justification for subsidiarity.

One of the limitations on centralized authority is the unwieldiness of surveilling and administering lower hierarchical levels from a centralized position. AI purports to solve this problem by making it easier for a single

[26] For the importance of participation, see David Hollenbach, *The Common Good and Christian Ethics*.

authority to use a single system to control a large organization. Information can be centrally gathered and analyzed, and then predictions implemented through a single algorithm. AI thus seems to make centralized authority easier and more efficient.

Finally, even if ultimate responsibility lies with more local levels of organization, they could still be using AI tools that are centrally designed. Because of the expense of developing software and AI from scratch, many small organizations will purchase AI from large corporations. For example, healthcare systems choose from a limited set of electronic medical records providers. Local governments buy software packages designed elsewhere. Such systems may be tailored to local organizational circumstances in some ways, but the broader tools and workflow will still be centrally designed. Thus, local action will be canalized and in part predetermined in its basic outlines by large corporations.

Therefore, subsidiarity and democratic participation can be undermined by algorithmic governance. The use of AI tends toward centralization, and even when used at local levels it can end up stereotyping action. Governance through AI remains opaque and unaccountable. All these features tend to undermine human agency.

In both areas discussed in this chapter, AI applications undermine the individual requirements for effective agency. In the workplace, deskilling removes the habits, skills, and virtues that are part of effective action. Algorithmic governance does a similar thing in the public sphere by eliminating opportunities for prudent deliberation by government officials and citizens in subsidiary organizations. It also prevents citizens from intersubjective engagement with those responsible for the decisions affecting them. In these ways, AI applications can remove possibilities for individuals to exercise agency.

CHAPTER 7

UNDERMINING AGENCY THROUGH MISINFORMATION AND RAPIDIFICATION

As chapter 2 discussed, responsible agency requires conscious deliberation. Wise deliberation requires both good information upon which to deliberate and time to slow down to deliberate. In part, ensuring these aspects are present depends on us as individuals. Individual virtues allow a person to pause, reflect and seek out guidance from others.[1] Yet they also depend on a human ecology because of the social influences on agency discussed in chapter 3. There are many cultural preconditions that enable access to knowledge and time for reflection. In some ways, AI could support these aspects of human deliberation. One of the claims made about generative AI is that it will allow access to and summaries of far more facts about a problem than we could find on our own. But that access is only valuable if these facts are true. Too often, generative AI serves the user hallucinations of its own devising or based on misinformation, a problem that may become worse as AI continues to train on text that will now include significant amounts of LLM output. AI boosters also promise us more time as AI takes over drudgery, allowing time for deliberation. However, many people feel as if the world is speeding up, allowing even less time for thought. This chapter explores these ecological factors of

[1] See discussion of the quasi-integral parts of prudence in Aquinas, *ST*, I-II, q. 49; Scherz, *Tomorrow's Troubles*, chapter 10.

misinformation and rapidification and how they affect responsible human agency.

Misinformation

Experts distinguish between "misinformation (false or misleading information without intent to deceive) and disinformation (false or misleading information intended to deceive)."[2] For convenience of writing, we will here largely just use the term misinformation to encompass both, unless we are specifically referencing disinformation. As a parasitic phenomenon, misinformation depends on some level of connection with factual information, yet it ultimately distorts reality by selectively omitting aspects, exaggerating others to the point of misrepresentation, and even fabricating entirely new facts. Misinformation is not new to human history. The Romans, for instance, embellished their military victories for propaganda purposes, as in Julius Caesar's portrayal of events in his *Commentaries on the Gallic War*.[3] Similarly, medieval chroniclers and modern historians have exaggerated or distorted historical events to serve political or ideological interests, making historical revision a necessity even centuries later. The intentional misrepresentation of facts is a recurring phenomenon, and, historically, misinformation could be corrected over time to a certain extent through scholarly inquiry and fact-checking.

Misinformation today is not just an amplified version of historical distortions, however. It is enhanced by new technologies that blur the lines between reality and fabrication. The sheer speed, scale, and sophistication with which information, and its opposites, spreads, is unprecedented: "Content can be relayed among users with no significant third-party filtering, fact-checking, or editorial judgment. An individual user with no track record or reputation can in some cases reach as many readers as Fox

[2] AI Research Group, *Encountering Artificial Intelligence*, 191. More broadly, this section draws on pp. 187–192 and 209–217 of that work.

[3] Michael Grant, *Greek and Roman Historians: Information and Misinformation* (Routledge, 1995).

News, CNN, or the *New York Times*."[4] Falsehoods now spread exponentially through social media platforms, bypassing accountability and jeopardizing agency and autonomy. This section reviews the role of knowledge in agency, before discussing how AI erodes truth in a way that undermines social cohesion.

Agency, Knowledge, and Truth

Effective agency requires an accurate grasp of reality even if, as human beings, we do not have a complete and perfect grasp of the world. While there are many theories of truth in philosophy, we regard truth as grounded in the way things are, encompassing both their present state and their potential to change. In some cases, reality is shaped by human action. Through our plans, projects, intentions, and desires, we modify reality to align with the designs that guide our actions. This kind of truth is known as "practical truth." Practical truth deals with aspects of reality that depend on us, and truth in practical reason is related to action, to how successfully we can change things to achieve our goals and designs. Many domains of human life are mediated by practical truth, including the arts, technology, and the social structures that define our institutions and roles, as well as language itself.

However, as we discussed regarding free will, to access and act upon practical truth, a person needs a well-rooted theoretical truth—namely, a good sense or grasp of those aspects of the world that do not depend on our action. The fact that triangles are three-sided figures, or that water and electricity do not mix well, are aspects of the world that we are unable to change. However, acquiring knowledge of properties of the world helps us succeed in practical actions, such as building and designing houses.

Aristotle recognized that truth is sought through both theoretical and practical reason, precisely because reality consists of both what exists

[4] Hunt Allcott, and Matthew Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives* 31, no. 2 (2017): 211–236.

independently of our goals and what falls within their reach.[5] This fundamental condition for truth has been articulated by some philosophers as the mind-to-world direction of fit, which certain mental states exhibit.[6] For example, for a belief to be true, our minds must align with the reality of what is happening in the world. In action, by contrast, the agent attempts to change the world to fit her intentions and desires. In misinformation, the direction of fit-for beliefs—mind-to-world—is absent, and in some cases, it is reversed into a world-to-mind direction of fit: Our fabrications aspire to become realities, and to be taken as facts to which our minds must conform.

We can categorize as departures from reality phenomena such as hallucinations, illusions, delusions, false beliefs, lies, and errors. These are all instances in which an accurate grasp of a mind-independent reality is absent. Such discrepancies may arise from cognitive malfunctions or intentional alterations of reality, whether for entertainment, through the creation of fictional worlds, or (in more troubling cases) due to deception or sheer ignorance.

However, it is important to distinguish between fiction and disinformation. Fiction, far from simply misrepresenting reality, often serves to illuminate aspects of the real world that are difficult to express directly. Crucially, creators of fiction do not claim that the worlds they construct correspond to actual reality. They signal this through various means, such as media of communication including novels, films, and other art forms. AI can assist people in creating fictional representations that help us better interpret the world. Without some indication of its fictional status, an expression intended as fiction could be mistaken for a representation of the real world, thereby reducing it to mere falsehood.

Not incidentally, many efforts to reduce misinformation in our digital era focus on making the indicators of its truth status more explicit. For

[5] Aristotle, *Nicomachean Ethics*, 1139b, 14-15.

[6] See G. E. M. Anscombe, *Intention*, 2nd ed. (Cornell University Press, 1963); J. R. Searle, *Intentionality: An Essay in the Philosophy of Mind* (Cambridge University Press, 1983).

instance, Twitter (now X) introduced labels for misleading tweets, and Facebook partnered with independent fact-checkers to flag false claims. AI-powered tools can also help detect and flag misinformation in real time, such as Google's Jigsaw project, which develops AI-driven fact-checking systems. Equally crucial is that misinformation often carries subtle indicators of its falsehood. For example, fake news tends to exhibit specific traits—such as lengthy titles, scant content, and frequent typos—that a media-literate public could identify and critically assess.[7] However, these efforts may ultimately prove insufficient if the sincerity conditions—that is, the speakers' commitment to the truth of what is stated, or to the fictional status of what is presented—are undermined by uncooperative agents who hide their communicative intentions or by algorithms that lack any kind of commitment to the truth or falsity of what they state, but merely calculate and predict outputs.

The Erosion of Language and the Inversion of Truth

In this context, it is no surprise that misinformation distorts the efficacy of human agency. This is particularly evident in how misinformation 1) erodes the integrity and function of language; and 2) alters the relationship between truth and reality, making truth dependent on individual or group desires and actions rather than on objective facts.

Misinformation constitutes an abuse of language, undermining one of its fundamental structures: its commitment to truth. Philosophers have emphasized that language, and in particular assertive speech acts, entails a commitment to the truth of what it asserts. Aquinas, for instance, located truth explicitly at the level of the judgment or the proposition, emphasizing that when we make a statement, we are not only saying

[7] Benjamin Horne and Sibel Adali, "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire Than Real News," *Proceedings of the International AAAI Conference on Web and Social Media* 11, no. 1 (May 3, 2017): 759–766, doi.org/10.1609/icwsm.v11i1.14976.

something but also committing ourselves to its truth.[8] This dual aspect of adequacy to reality and reflexivity, namely of committing to the truth of what is said, marks the nature of truth in language. Similarly, in speech act theory, J. L. Austin introduced the concept of felicity conditions,[9] which determine whether a statement successfully performs its intended communicative function. Paul Grice's "maxim of quality" further reinforces this idea by stating that speakers should not assert what they believe to be false and should not say something for which there is not adequate evidence. Since communication depends on a shared commitment to truthfulness, one of Grice's cooperative principles, which governs effective communication, reads "try to make your contribution one that is true."[10]

Without this commitment, the structure of language collapses. If someone says, "Paris is the capital of France," the utterance presupposes that this statement corresponds to reality. However, if speakers systematically use language without such an ontological commitment (whether through ignorance, distortion, or deliberate ambiguity) then expressions cease to carry fixed meanings. If "Paris is the capital of France" could arbitrarily mean the capital of Italy, Nigeria, or Japan, or even "Let's not talk about Paris at all," language would lose one of its uses: its function as a vehicle for conveying certain truths about aspects of the world. This would lead to a communicative breakdown. Analytic philosophers refer to this fundamental commitment to truth in language as "sincerity conditions."[11]

As noted, misinformation also distorts the basic structure of truth, which conforms to reality (mind-to-world direction of fit).[12] Instead,

---

[8] Thomas Aquinas, *Commentary on the Metaphysics of Aristotle*, Lib. 6, lect. 4.

[9] J. L. Austin, *How to Do Things with Words* (Harvard University Press, 1975).

[10] Paul Grice, "Logic and Conversation," in *The Logic of Grammar*, ed. D. Davidson and G. Harman (Dickenson, 1975), 64–75.

[11] J. R. Searle, *Speech Acts: An Essay in the Philosophy of Language* (Cambridge University Press, 1969).

[12] J. R. Searle, "What Is an Intentional State?" *Mind* 88, no. 349 (1979): 78.

misinformation inverts this relationship, imposing a world-to-mind direction of fit, where reality is expected to conform to human desires, biases, or ideological agendas, rather than being acknowledged as it is. This reversal of truth has profound consequences. When truth is subordinated to subjective interests, action itself loses its grounding. Efficacious agency depends on an accurate understanding of the world; without it, actions become misguided, ineffective, or even harmful. If decisions and policies are based on fabricated or distorted information, individuals and societies lose their ability to navigate reality effectively.

Paradoxically, by attempting to reshape reality to fit human preferences rather than adjusting beliefs to match reality, misinformation weakens rather than enhances agency. The distortion of facts, whether intentional or unintentional, impedes a reliable grasp of reality, undermining both personal decision-making and collective progress. Without a stable foundation of truth, meaningful action becomes impossible, as it lacks the necessary orientation toward the real world. As John Paul II states, "Truth enlightens man's intelligence and shapes his freedom, leading him to know and love the Lord."[13]

The Weakening of Social Cohesion

Ultimately, the degradation of language and the distortion of truth weaken the social bonds that underpin a healthy society. The sincerity condition inherent in communication is a foundational element of social life. If language's natural commitment to truth is compromised, trust among speakers erodes, undermining cooperation and mutual understanding.

Both forms of erosion, the corruption of language and the subordination of truth to subjective interests, have contributed to a widespread decline in trust of various institutions, from governments to financial markets, from scientific bodies to public health organizations. This distrust is particularly evident in crisis situations, where misinformation spreads rapidly, exacerbating confusion and obstructing

[13] John Paul II, introductory blessing to *Veritatis Splendor*.

effective responses. False information about health crises, natural disasters, and conflicts not only misleads individuals but also hinders coordinated efforts to address urgent challenges.

When misinformation becomes pervasive, it not only corrupts individual reasoning but also dismantles the structural integrity of democratic institutions and social cooperation.[14] A society that cannot agree on basic facts loses its ability to make informed decisions, leading to governance failures, public distrust, and societal fragmentation. In such an environment, individuals are more susceptible to manipulation, and collective action becomes increasingly difficult.

If the social bond is weakened by the intentional or unintentional misuse of language, societal cohesion no longer emerges organically from individuals freely engaging in the pursuit of the common good, something that institutions and social norms also foster. Instead of a democratic society in which free agents choose to pursue shared goals and contribute to collective wellbeing, misinformation can lead to social cohesion being imposed externally, either by a small, powerful minority that controls the dominant narrative or by impersonal, mechanized algorithms that exploit human biases, fears, and the drive for monetization.

In the first case, misinformation is deliberately crafted and strategically deployed to manipulate public opinion, ensuring submission and consolidating control. This phenomenon is particularly evident in disinformation campaigns, in which state and/or non-state actors intentionally spread falsehoods to manipulate public perception, influence elections, and sow discord. In such cases, individuals are not encouraged to engage in critical reflection or meaningful discourse; instead, inflammatory language and images condition them to react impulsively, driven by fear, desire, or immediate gratification. This manipulation of the public mind is particularly dangerous because it shifts the basis of collective decision-making away from truth and reason,

[14] Paul Scherz and Luis Vera, "AI and the Subjective Crisis of Knowledge," *Journal of Religious Ethics* (July 2025), doi.org/10.1111/jore.70001.

replacing it with a landscape where fabricated narratives, emotional appeals, and ideological distortions dictate social and political action. As rational deliberation is undermined, public discourse becomes fragmented, polarization intensifies, and the ability to engage in meaningful, consensus-driven dialogue is severely diminished.

Digitalization also enables the creation of AI-generated text and synthetic media that increasingly complicates the distinction between real and forged content. Deepfake technology, which uses AI to manipulate video and audio, has been used for both entertainment and malicious purposes, such as falsely depicting political figures saying things they never did. For instance, a deepfake video of Ukrainian President Volodymyr Zelenskyy supposedly surrendering to Russia circulated online in 2022, suggesting the potential of AI-generated misinformation to influence public opinion.[15] Similarly, scammers can exploit voice cloning technology to impersonate family members in distress, leading to financial fraud.

Beyond intentional fabrications, algorithmic amplification is another major contributor to misinformation. In this case, misinformation operates through impersonal, algorithmic mechanisms that exert a profound influence on society. Digital platforms, fueled by engagement-driven algorithms, prioritize sensationalism, controversy, and emotionally charged content, often at the expense of truth. Recommendation algorithms, designed to maximize user engagement, contribute to this effect by selectively exposing users to information that confirms their biases and the digital structures that reinforce them, creating a feedback loop that makes countering falsehoods increasingly difficult.

When truth is sidelined in favor of virality, individuals find themselves trapped in echo chambers, digital environments where people are exposed primarily to information that aligns with their pre-existing beliefs. Echo chambers reinforce biases, rather than critically evaluating competing

[15] Bobby Chesney and Danielle Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review* 107 (2019): 1753.

viewpoints, and thereby exacerbate societal polarization.[16] In fact, psychological research has shown that when confronted with information that contradicts their beliefs, people often minimize cognitive dissonance by interpreting facts through the lens of their existing worldview or rejecting the new information entirely. This phenomenon, known as the backfire effect, suggests that mere exposure to factual corrections does not necessarily change minds; sometimes, it even strengthens the original false belief.[17] AI-driven bots further complicate this issue by engaging in coordinated inauthentic behavior on social media, making it increasingly difficult to detect and counteract misinformation.[18]

Moreover, the monetization of information aligned with surveillance capitalism encourages the spread of false and misleading content because it attracts large audiences and generates significant ad revenue, incentivizing bad actors to produce and spread misinformation deliberately. Misinformation is often more engaging than factual reporting, meaning that platforms have a financial interest in keeping it circulating.[19] For example, during the 2016 US presidential election, fabricated stories outperformed real news in terms of engagement on social media.[20]

Last, with the advent of sophisticated artificial intelligence, misinformation has acquired an autonomous dimension. Initially designed for specific purposes, AI-driven content generation tools can quickly produce output that goes beyond the control of their creators. Large language models (LLMs) have generated false or misleading content,

---

[16] Cass R. Sunstein, *#Republic: Divided Democracy in the Age of Social Media* (Princeton University Press, 2017).

[17] Brendan Nyhan and Jason Reifler, "When Corrections Fail: The Persistence of Political Misperceptions," *Political Behavior* 32, no. 2 (June 2010): 303–330, doi.org/10.1007/s11109-010-9112-2.

[18] Emilio Ferrara, "The History of Digital Spam," *Communications of the ACM* 62, no. 8 (2019): 82–91, doi.org/10.1145/3299768.

[19] Soroush Vosoughi, Deb Roy, and Sinan Aral, "The Spread of True and False News Online," *Science* 359, no. 6380 (March 9, 2018): 1146–1151, doi.org/10.1126/science.aap9559.

[20] Allcott and Gentzkow, "Social Media and Fake News in the 2016 Election."

as seen in the case of ChatGPT fabricating sources when asked for academic references or even legal precedents.[21] Moreover, although neutrally designed according to stochastic models, "AI can be maliciously led to learn biased models," namely by "deliberately injecting spam—here intended as unwanted information—into the training data," all while portraying the appearance of neutrality.[22]

These unintended effects of information technology have made obvious the need for algorithmic transparency. Some have argued for mitigating misinformation risks "by budgeting for curation and documentation at the start of a project and only creating data sets as large as can be sufficiently documented."[23] Decision-making systems, especially in areas like medical diagnosis, loan approvals, autonomous vehicle safety, and AI-powered hiring tools, should provide human-readable explanations of their decisions. Algorithmic transparency hopes to ensure fairness, accountability, and trust in AI systems. However, deep learning models are highly complex, making full transparency difficult. Moreover, while independent audits and documentation can help improve accountability, disclosing too much about an algorithm's inner workings may leave platforms vulnerable to exploitation.[24] Explainability must therefore be balanced with privacy, security, and intellectual property concerns.

Given the vast amount of data that an AI can process as compared to human cognitive capacities, along with their success in many tasks, it is easy to overestimate their reliability and place undue trust in their performance.

---

[21] Emily M. Bender, 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* (ACM, 2021), 610–623, doi.org/10.1145/3442188.3445922.

[22] Emilio Ferrara, "The History of Digital Spam," *Communications of the ACM* 62, no. 8 (July 24, 2019): 82–91, doi.org/10.1145/3299768.

[23] Bender et al., "On the Dangers of Stochastic Parrots," 610.

[24] Zachary C. Lipton, "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery," *Queue* 16, no. 3 (June 2018): 31–57, doi.org/10.1145/3236386.3241340.

For example, "LLMs are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form." Focusing on state-of-the-art results "without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results."[25] LLMs merely manipulate linguistic form without understanding meaning or committing to the truth of their output, which raises the question of whether they deserve the same degree of credibility as human speakers, who use language meaningfully and with a sincerity condition. LLMs lack sincerity conditions because they do not engage in genuine belief or truth commitment; they simply generate responses based on statistical patterns and predefined parameters for coherence and correctness.

The growing sophistication of misinformation presents significant challenges for society, from eroding trust in institutions to undermining individual agency. Legislators, educators, researchers, and tech companies have scrambled to implement solutions, often struggling to keep pace with the evolving landscape, much like a parent chasing after a toddler who has just learned to walk and attempts to run into traffic. However, through a combination of technological solutions, regulatory oversight, and digital literacy initiatives, there is potential to mitigate at least some of its impact. For example, decentralized technologies like blockchain offer a way to combat misinformation and fraud. Blockchain's cryptographic security and transparent ledger system can be used to authenticate digital content, preventing fraud and ensuring accountability, provided that the right infrastructures for authenticating information are in place. Mechanisms that allow for a higher level of verification are particularly relevant for digital voting systems, non-fungible tokens (NFTs), digital currencies, and

[25] Bender et al., "On the Dangers of Stochastic Parrots," 610.

even contract management, where ensuring the authenticity of information is critical.[26]

Legislation can also help address these problems. For example, on November 16th, 2022, the European Union began to enforce the Digital Services Act (DSA), which holds major platforms like Meta, X, TikTok, and Google responsible for mitigating the spread of harmful or false information. Platforms are required to remove misinformation quickly and share algorithmic data with regulators and researchers, among other regulations. In the US FOSTA-SESTA (Allow States and Victims to Fight Online Sex Trafficking Act and Stop Enabling Sex Traffickers Act) amended Section 230 of the Communications Decency Act (1996) and passed it into law in 2018, allowing for civil and criminal prosecution of platforms if they knowingly facilitate, assist, or promote sex trafficking or prostitution of others. However, other proposals are still under consideration or have been introduced but not yet enacted. Balancing innovation with accountability will be key in shaping an information ecosystem where truth, rather than deception, remains the foundation of collective knowledge and decision-making.

Restoring trust in language and truth requires reinforcing the mechanisms that promote epistemic responsibility, both at the level of individuals and institutions. Encouraging critical thinking and fostering media literacy are essential steps toward rebuilding a society where truth, rather than misinformation, serves as the foundation for meaningful discourse and collective decision-making.

Beyond any countermeasures that governments and digital platforms may devise, nothing substitutes for the personal responsibility of each individual to uphold the fabric of society through a truthful use of language, which permeates and mediates our social world. In fact, freedom of speech is also at risk when misinformation prevails and governments

[26] Horst Treiblmaier, "Combining Blockchain Technology and the Physical Internet to Achieve Triple Bottom Line Sustainability: A Comprehensive Research Agenda for Modern Logistics and Supply Chain Management," *Logistics* 3, no. 1 (2019): 10.

resort to draconian measures to minimize it, potentially limiting our right to speak freely. Being part of society and coexisting with others requires a respectful use of language. Such respect must be offered both toward the structure of language itself, which by default and through its very use commits us to the truth of what is said, and toward the fact that language, as a shared and cooperative endeavor among speakers, derives its rules and meaning from its public use.[27]

Misinformation and Agentic Systems

The subversion of the news that people gather on the Internet and social media is the most discussed, but not the only, way that malicious actors threaten our information ecosystem. There can also be disinformation targeted at agentic systems themselves. Of particular concern are attacks on public sources of code, especially given the widespread use of LLMs as software coding agents. These agents can draw portions of code from public sources such as GitHub. In what is known as a watering-hole attack, bad actors craft packages of code containing a Trojan horse or other malicious software. They give these common names (or names commonly hallucinated by LLMs) and wait for these packages to be incorporated into others' code. Since AI-generated code is notoriously difficult for human programmers to read or understand, the programmer is unlikely to spot the malicious directions. Malicious prompts can be easily hidden. One simple method is to write in white text on a white background, producing text that will go unnoticed by humans but will still be read and executed by an LLM.

When AI agents are used to perform actions automatically, without user intervention, risk is accelerated precisely because the agents have tunnel vision. AIs fail to see a larger picture that might tell them to check certain pieces of code or avoid certain actions. They cannot distinguish between what should be trusted and what should not. And prompts such as "do only what is legal" or "download only safe code" are meaningless.

[27] Ludwig Wittgenstein, *Philosophical Investigations* (1957).

Adir Gruss, cofounder and CTO of Aim Security, believes these issues will ultimately require a fundamental redesign of AI agents: "Imagine a person that does everything he reads—he would be very easy to manipulate. Fixing this problem would require either ad hoc controls, or a new design allowing for clearer separation between trusted instructions and untrusted data."[28] As AI expert Gary Marcus notes, "As long as we have agents roaming the internet and otherwise incorporating data that they don't fully understand—and LLMs don't ever fully understand the data they are leveraging—there is enormous risk."[29]

These problems with coding are only one of the significant security issues raised by AI programs that act as agents on a user's behalf. AI agents used to book travel, make purchases, or handle legal matters add a middleman to our transactions, giving malicious actors another point of entry from which to acquire sensitive information such as credit cards or government identification numbers. In an organization, the use of AI agents risks the leaking of private or sensitive information. It also risks exposing an entire system to data corruption. Hallucinations, still common to LLMs, can easily corrupt data and have cascading effects. To work effectively, agents often require users' passwords, yet in a recent trial of current AI programs, all failed to handle passwords in a consistently secure manner, thus opening the door to a systemic attack.[30] While AI-driven misinformation thus threatens social stability, it is only one of the threats that new forms of AI are bringing to society.

---

[28] Sharon Goldman, "Exclusive: New Microsoft Copilot Flaw Signals Broader Risk of AI Agents Being Hacked—'I Would Be Terrified,'" *Fortune*, June 11, 2025, fortune.com/2025/06/11/microsoft-copilot-vulnerability-ai-agents-echoleak-hacking.

[29] Gary Marcus and Nathan Hamiel, "LLMs + Coding Agents = Security Nightmare," *Substack*, August 17, 2025, garymarcus.substack.com/p/llms-coding-agents-security-nightmare.

[30] Maxime Rossi Bellom and Ramtine Tofighi Shirazi, "Is Vibe Coding a Security Nightmare? A Benchmark of AI Coding Agents," *SecMate Blog*, August 7, 2025, blog.secmate.dev/posts/vibe-coding-security-benchmark.

Rapidification

AI technologies promise increased efficiency and freedom from mundane tasks. Yet, paradoxically, as more tasks become automated through technology, users are pressured to increase their productivity. Sociologists such as Hartmut Rosa have referred to this phenomenon as "social acceleration," characterized by a "time-famine" or sense of diminishing time to complete tasks.[31] Many of us witness this through our inboxes: We receive many more emails throughout the day than we would have received letters one hundred (or even forty) years ago. Apps that make it possible to complete administrative tasks that might otherwise have required specialized expertise reshape the job market, placing the burden on each of us to accomplish more demands ourselves. Similarly, social media and the Internet make us more aware of tasks to accomplish, products to investigate, or sources to peruse. Rather than expanding our free time to use at our discretion, technology often leads to expectations and pressures on our time by forces of which we otherwise might not have been aware, placing increased burdens on our agency.

Thus, those of us who live in Western cultures seem to be doing ever more things in less time while often feeling under pressure to do more. Rosa describes this as a vicious cycle in which new technologies increase the rate of social change, which in turn speeds up our daily lives, leading to a search for further time-saving technologies. The "mythical logic of incessant growth, innovation and escalation that permeates all spheres of life . . . imposes itself on everybody, independently of their will. The idea that one should always maintain one's options open and expand the reaches of one's action has transformed life into a perpetual rat race."[32] Hence we begin to perceive the Darwinian struggle of the fittest as the

[31] Hartmut Rosa, *Social Acceleration: A New Theory of Modernity* (Columbia University Press, 2013).

[32] Frederic Vandenberghe, "Tuning into Harmut Rosa's Systematic Romanticism," *Journal of Chinese Sociology* 10, no. 12 (2023): 18.

"survival of the fastest."[33] Pope Francis referred to this phenomenon as "rapidification." In *Laudato Si'*, he observed that "Although change is part of the working of complex systems, the speed with which human activity has developed contrasts with the naturally slow pace of biological evolution. Moreover, the goals of this rapid or constant change are not necessarily geared to the common good or to integral and sustainable human development."[34] To this point, many of the defenses of AI speed involve examples of destructive competition: the need to outmaneuver an enemy on the battlefield, to develop a software program before others, or to take advantage of opportunities for arbitrage before other traders.

As our society speeds up, we attempt to do more and more by multitasking. The average office worker is interrupted roughly every three minutes during the workday, interruptions that the worker herself often initiates, such as by checking email or social media.[35] As more jobs move to the knowledge and service sectors, more of us find the temptation to multitask at work irresistible. However, studies have shown that, contrary to self-perception, the human brain cannot concentrate on several things simultaneously. Rather, it switches focus from task to task. Thus, all tasks are slowed down by the switching process, and the attempt to coordinate too many tasks can lead to more time devoted to switching than to the tasks themselves, a phenomenon that computer scientists have long known as "thrashing." Over time, multitasking erodes our ability to pay focused, close attention, and this eventually eats away at traits such as patience, tenacity, judgment, and the ability to problem-solve.[36]

---

[33] Hartmut Rosa and William Scheuerman, eds., *High-Speed Society: Social Acceleration, Power, and Modernity* (Pennsylvania State University Press, 2009), 8.

[34] Francis, *Laudato Si'*, §18.

[35] Victor M. González and Gloria Mark, "'Constant, Constant, Multi-Tasking Craziness': Managing Multiple Working Spheres," *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM, 2004), 113–120, doi.org/10.1145/985692.985707.

[36] "Multitasking: Switching Costs," *American Psychological Association*, apa.org/topics/research/multitasking.

If multitasking is so deleterious to both our productivity and our character, why do we do it? Multitasking and the need to do everything faster are responses to a culture of "more" and the demands it places upon us. They are also vain attempts to escape from our own finitude. Rosa describes social acceleration as a modern attempt to attain the fullness of eternal life within the boundaries of mortality.[37] Reinhold Niebuhr identified our inability to accept the natural limits on our time and abilities as a source of sin: "Man is ignorant and involved in the limitations of a finite mind; but he pretends that he is not limited. He assumes that he can gradually transcend finite limitations until his mind becomes identical with the universal mind. All of his intellectual and cultural pursuits, therefore, become infected with the sin of pride."[38] In our pride, we pursue unlimited knowledge, accomplishment, and experience rather than accepting our human finitude and relaxing into simply being what we are.

Commitment to efficiency and speed cannot ensure that all human goods are met. Some goods require time to mature and develop—and even may require our slow growth and consistent encounter to properly appreciate. In fact, the highest goods are meant to be savored and appreciated. These include establishing authentic and deep relationships with each other, with God, and with the world around us.

AI can distract us from direct engagement, either by providing a filter through which we engage the other or by acting as a substitute for human engagement. Generative AI programs offer an immense load of information with no assurance of veracity. They provide answers in a flash, yet, as we saw in the last section, current generative AI programs are not designed to distinguish between truth or falsehood. While we may have extensive information at our fingertips, we often lack the time or

[37] Rosa, *Social Acceleration*, 180–182. For theological analysis, see Paul Scherz, "Living Indefinitely and Living Fully: *Laudato Si'* and the Value of the Present in Christian, Stoic, and Transhumanist Temporalities," *Theological Studies* 79, no. 2 (2018): 356–375.

[38] Reinhold Niebuhr, *The Nature and Destiny of Man: A Christian Interpretation* (Westminster John Knox, 1996), 2:178–179.

motivation to assimilate and evaluate the truth or falsity of that information. Valuing immediate results and fast insights over truth, we undervalue the need for conversation, counsel, and discernment.[39] As *Antiqua et Nova* acknowledges, "AI cannot currently replicate moral discernment or the ability to establish authentic relationships. Moreover, human intelligence is situated within a personally lived history of intellectual and moral formation."[40] Access to convenient search results, often biased by commercial forces, renders us less likely to conduct a proper investigation into all options available to us. Thus, trust in Internet searches or AI feedback isolates the user, obviating the insights that might have been gained from the tested life experience of our neighbor.

Close, focused attention is essential to our communal life. Giving another our fullest attention is, according to Simone Weil, "the rarest and purest form of generosity."[41] Fully attending to another is something that cannot be rushed. Buddhist scholar Alan Wallace notes: "When we give another person our attention, we've given away that portion of our life. We don't get it back. We're giving our attention to what seems worthy of our life from moment to moment. Attention, the cultivation of attention, is absolutely core" to being with one another.[42] Jesus calls it the greatest love to give up one's life for one's friends. While we think of that in literal terms, Wallace points to the fact that in giving our time and attention to someone we do give a part of our life.

We are also called to slow down to grow spiritually. John Paul II wrote in *Laborem Exercens*, "Therefore man's work . . . cannot consist in the mere

[39] Cf. Mariele Courtois, "Among His Own Kin and in His Own House: Artificial Intelligence and the Interrelational Cultivation of Prudence," *Journal of Religious Ethics*, Special Issue (2025): 1–21.

[40] Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §32.

[41] Simone Weil, Letter to Joë Bousquet, April 13, 1942, in Simone Pétrement, *Simone Weil: A Life*, trans. Raymond Rosenthal (Pantheon, 1976).

[42] Maggie Jackson, *Distracted: The Erosion of Attention and the Coming Dark Age* (Prometheus, 2008), 259.

exercise of human strength in external action; it must leave room for man to prepare himself, by becoming more and more what in the will of God he ought to be, for the 'rest' that the Lord reserves for his servants and friends."[43] Rest allows us to take part in the greatest form of work: the work that involves not simply service for material needs but also the stillness at the service of our spiritual needs and commitment to our eternal, immoveable call from God to his lasting refuge.[44] Leo XIII writes: "The rest from labor is not to be understood as mere giving way to idleness; much less must it be an occasion for spending money and for vicious indulgence, as many would have it to be; but it should be rest from labor, hallowed by religion."[45] God is the ultimate exemplar of rest, which God relished following the work of creating the universe.

One discipline that eschews speed is the Benedictine practice of *lectio divina*, or meditative reading, in which the practitioner focuses on a short passage of scripture or another work, reading it slowly and repetitively. Taking time to let the words sink in, to connect them with the reader's own experience and self, is the only way to receive its benefits. It is one example of what Harris Wiseman calls "slow knowing": "Knowing slowly implies a kind of patience, a willingness to engage with something, to give things more than a merely cursory glance or scan, and a willingness to engage with things in a way that gives them the time that is needed for deeper layers of meaning to unfold."[46] While we can treat religious propositions as mere facts, stories, or something to be intellectually analyzed, slowly sitting with them can turn them into objects of care. There is a difference between taking in text as data, giving intellectual assent, and letting it form one's character and life. Further, our own social communications have a vast and immediately impactful influence, and our

[43] John Paul II, *Laborem Exercens*, §25.
[44] Psalm 91:4.
[45] Leo XIII, *Rerum Novarum*, §41.
[46] Harris Wiseman, "Knowing Slowly: Unfolding the Depths of Meaning," *Zygon* 57, no. 3 (September 2022): 721–722.

cooperation with these structures should involve respect for the persons whose consciences and imaginations are impacted by the information shared.

Beyond intellectual comprehension, we might also ponder the text as a kind of prayer which, ideally, leads into a state of contemplation.[47] Spiritual contemplation takes time. It relaxes our preoccupation with the onrushing arrow of time, introducing a vertical dimension. Psalm 46:11 counsels us to "Be still and know that I am God!" Indeed, one of the goals of meditation is precisely to slow down the frenetic activity of the mind. Can AI aid us in contemplation? AI agents can take over some of our tasks—collecting data, booking trips and appointments, drafting routine emails. This might indeed give us more time. But if we spend that time simply adding to our to-do lists, we have gained little. Niebuhr reminds us never to lose sight of our nature as finite creatures. Our tools might make us more efficient in our work of caring for nature and for each other. Yet our deepest need, as we do the work at hand, is to give our full attention to God and to one another. For it is those relationships that are ultimately the goal of all our doing.

AI and social media, through both passive consumption and active participation, foster habits of speed over habits of patience. Algorithms are deliberately written to promote anger and division (which increase emotion and therefore "engagement") rather than thoughtfulness or peace of mind. Spiritual growth demands discretion rather than immediate gratification. Benedict XVI writes that considering technologies' "fundamental importance in engineering changes in attitude towards reality and the human person, we must reflect carefully on their influence, especially in regard to the ethical-cultural dimension of globalization and the development of peoples in solidarity."[48]

---

[47] Harris Wiseman, "Language Use in Spiritual Practice," in *Perspectives on Spiritual Intelligence*, ed. Marius Dorobantu and Fraser Watts (Routledge, 2024), 167.
[48] Benedict XVI, *Caritas in Veritate*, §73.

AI may soon seem to provide a bespoke reality for each user, one that immerses us in constant activity and gratification. Nature, which unfolds at speeds almost imperceptible to us, could come to seem dull and uninteresting. Yet nature itself is, as Pope Francis notes, "a magnificent book in which God speaks to us and grants us a glimpse of his infinite beauty and goodness. . . . Rather than a problem to be solved, the world is a joyful mystery to be contemplated with gladness and praise."[49] To celebrate the wonders of God evident in creation requires careful, trained attention to the world around us. We must allow our surroundings to unfold signs of God's presence at their own pace.

So long as our technology needs to operate interactively with natural and human agents, it is bounded by the upper limit of our capabilities. As Dario Amodei notes, "Animals run at a fixed speed so experiments on them take a certain amount of time which may be irreducible. The same is true of . . . anything involving communicating with people."[50] We can only move so fast. However, a fully autonomous AI can move much faster, at speeds at which humans cannot keep up. We saw an example of this in the 2010 "flash crash," when automated trading programs began to trade at a pace and volume that human traders on the floor were unable to match or even follow.

In no realm of human action is the interplay between speed and autonomy of more concern than in the deployment of lethal autonomous weapons. While most commanders express a desire for a human to be "in the loop" of military decision-making, how much control can they have if decisions in the field are made at speeds humans are unable to follow? The tempo of war has steadily accelerated over time. At some point, as reliance on AI expands, people might be forced to cede all decision-making to machines. This is particularly salient should we begin to rely on AI to make strategic rather than merely tactical decisions and could eviscerate any

[49] Francis, *Laudato Si'*, §12.

[50] Dario Amodei, "Machines of Loving Grace: How AI Could Transform the World for the Better," October 2024, darioamodei.com/essay/machines-of-loving-grace.

meaning from the concept of "mission command." However, "Human dignity depends on our working with our tools rather than letting them supplant us and this is at its most important in matters that involve questions of life and death."[51]

Science fiction has explored concerns, now echoed by numerous scientists, that we might lose our agency to an AGI or superintelligent computer. We want our digital creations to remain under human control, even when they have powers and abilities that surpass our own. For this to remain the case, we must ensure continued interaction between AI and humans, which is only possible when AIs operate at some level at human speed. Similarly, as we race to produce ever more intelligent AIs, we might consider that the Silicon Valley motto to "move fast and break things" is neither wise nor good. For the things we may break may be ourselves, our societies, our planet, and our faith, in each other and in God.

[51] Noreen Herzfeld and Robert Latiff, "Can Lethal Autonomous Weapons Be Just?" *Peace Review: A Journal of Social Justice* 33 (2021): 218.

CHAPTER 8

FUTURE POSSIBILITIES FOR AI AND AGENCY

The main question of this book is whether people will design and use AI in ways that enable or diminish human freedom. As we move into the future, this question will likely become more pressing. Our approach to issues at the intersection of AI and human agency must likewise be sharpened—both in response to new developments in AI and through careful reflection on how a Catholic approach to AI aligns with and departs from rival views of the future of humanity.

Consider, for example, a transhumanist approach. Transhumanists aim to eclipse limitations on humanity through technological means. While transhumanists regularly caution against the existential threats posed by AI, they are overall optimistic about future developments in AI and their contribution to human flourishing.[1] This is because:

> Transhumanists view human nature as a work-in-progress, a half-baked beginning that we can learn to remold in desirable ways. Current humanity need not be the endpoint of evolution. Transhumanists hope that by responsible use of science, technology, and other rational means we shall eventually manage to become posthuman, beings with vastly greater capacities than present human beings have.[2]

[1] See, e.g., Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press), 2003.

[2] Nick Bostrom, "Transhumanist Values," *Journal of Philosophical Research* 30, supplement (2003): 4.

In its optimism about the future and the potential of new technologies to achieve it, transhumanist anthropology betrays a commitment to what Pope Paul VI might call a purely material hope, one "based on a materialistic and atheistic philosophy, namely one which shows no respect for a religious outlook on life, for freedom or human dignity."[3] The transhumanist outlook also manifests the technocratic paradigm that Pope Francis cautioned against, a view according to which "artificial intelligence and the latest technological innovations start with the notion of a human being with no limits, whose abilities and possibilities can be infinitely expanded thanks to technology."[4] A Catholic approach to future developments in AI should not be a backward-looking rejection,[5] but also cannot share transhumanists' purely materialistic optimism.[6] Yes, Catholics, like transhumanists, should be careful of the existential threats posed by AI and other new technologies.[7] But when focusing on future possibilities, Catholics must also be cautious of adopting a purely material optimism. Instead, we are called to Christian hopefulness. What Christian hopefulness looks like, and how it measures up against transhumanist optimism, is our next topic.

---

[3] Paul VI, *Populorum Progressio*, March 26, 1967, §39, vatican.va/content/paul-vi/en/encyclicals/documents/hf_p-vi_enc_26031967_populorum.html.

[4] Francis, "*Laudate Deum*: Apostolic Exhortation to All People of Good Will on the Climate Crisis," October 4, 2023, §21, vatican.va/content/francesco/en/apost_exhortations/documents/20231004-laudate-deum.html.

[5] Brian Patrick Green, "The Catholic Church and Technological Progress: Past, Present, and Future," *Religions* (2017): mdpi.com/2077-1444/8/6/106.

[6] Brian Patrick Green, "A Roman Catholic View: Technological Progress? Yes. Transhumanism? No," in *Religious Transhumanism and Its Critics*, ed. Arvin M. Gouw, Brian Patrick Green, and Ted Peters (Lexington Books, 2022); Brian Patrick Green, "Transhumanism and Roman Catholicism: Imagined and Real Tensions," *Theology and Science* 13 (2015): 187–201.

[7] The Church seeks not only to remind everyone of the duty to care for nature in its work, but at the same time "she must above all protect mankind from self-destruction" (Francis, *Laudato Si'*, §79, referencing Benedict XVI, *Caritas in Veritate*).

Human Freedom and the Future

Consider a jarring question: Will humans be able to make decisions *at all* in the future? The question is not as far-fetched as we would like it to be. Already, as we discussed regarding deskilling, humans have ceded large swathes of our agency to AIs: People use LLMs to craft prose, navigation apps to explore new cities, and algorithms to suggest our entertainment choices. None of these practices demolish human agency, but each chips away at it. And with each loss we become increasingly dependent on AI, allowing it, or, in some cases, those humans who are developing and guiding it, to take on the role of a caretaker. But of course, AI is not a caretaker. It is a complex system that humans have created. Likewise, AI developers, even if well-intentioned, cannot provide the nurture and love required of caretakers. The risk, then, is not merely the dwindling of agency to the point of infantilization. As we look to the future of AI, we face the risk of becoming not merely infants but infants of neglectful parents. Less poetically, as AI is integrated into more areas of our lives, we must take care to guard our agency, lest it is eroded beyond recognition and we find ourselves unable to make decisions at all.

While the atrophying of human agency raises deeply concerning issues, in a Christian context answering questions about the future of decision-making is not enough. This is because Christians have a richer sense of freedom than simply our ability to make decisions. For the Christian, God enables human freedom, so already the idea of freedom as *autonomous* decision-making is challenged. Christians must therefore be concerned with questions not only about the erosion of human agency but also about how the future of AI will affect human freedom in a distinctively theological context where it is seen as directed toward conforming to the divine will.

For example, humans are free to break God's moral law. While nature has physical laws that cannot be broken (e.g., what goes up must come down), God's moral law can be rejected. But breaking these laws catches up with each individual eventually, whether in this life or the next. In

short, the mere exercise of freedom is not always good from a Christian perspective; it is right only when it is used to pursue good and avoid evil. So, from a theological perspective, it is not enough merely to ask *whether* we will be able to make decisions in the future. We must also ask, more specifically, whether AI will provide the kind of freedom that is good for us. As in the earlier sections on nudging, we can, for example, imagine future AI applications that limit human agency, but in doing so, increase the likelihood that we will do good and avoid evil; but we can also imagine the opposite.

As Christians navigate the future of AI and human agency, Christians cannot reduce these conversations to questions about decision-making. They must rather contextualize questions about human freedom in the theological truth that human freedom is grounded in our freedom in Christ, and the moral truth that freedom ought to be used for the good, and not for evil.

AI and The Future of Human Agency

AI has great capacity to be used for good.[8] As we have seen in the previous chapters, however, human agency is challenged by AI in myriad ways: through nudging, deskilling, algorithmic governance, disinformation, and the rapidification of modern life. As we look to the future, these challenges will only increase in scope and scale. Indeed, we are very likely to witness the development of faster AI with more computing power, more massive AI enabled through larger data sets, new software architectures and algorithms, and the implementation of quantum computing. These developments will, in turn, facilitate the integration and influence of AI in our lives. It will become increasingly possible for us to take human trends in power-seeking through technology toward their maximal ends: utter efficiency at getting what we want, complete scope for satisfying all desires, and perfect certainty that we can attain them. These trends, if properly

[8] Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §2.

directed, can be used for seeming goods: efficiency, the satisfaction of desires, increased confidence in attaining our goals. None of these aims are intrinsically disordered, and if we use AI to achieve well-ordered desires and proper ends, that can be a good thing. Yet not all desires are well-ordered, nor are all ends good. As AI is integrated into our lives, we must therefore be keenly aware of the ways that it can influence, undermine, or even thwart human agency.

Clearly, we cannot speculate about all the ways future developments of AI may challenge human agency. The examples we offer should therefore be considered not only speculative, but also non-exhaustive. But we can offer a framework for thinking about future challenges posed to human agency, an "inverted pyramid" schema according to which we can reflect on future challenges across multiple levels,[9] from cultural-environmental at the broadest to the integration of AI with human biology at the most intimate. This schema, we believe, can provide structure for reflecting on other ways AI may challenge human agency in the future beyond the specific examples we cover here.

At the broadest level, then, future developments in AI will challenge human agency by altering the social, cultural, and environmental conditions in which we manifest our agency, as with the rapidification and changed information ecology discussed in the last chapter. These challenges will often erode human agency in subtle and diffuse ways, yet in their breadth and pervasiveness, they have the potential to be particularly nefarious. For example, as vehicles become increasingly automated, transportation may become safer, but at the same time, our choices about whether and how to move through our environments will become increasingly restricted.[10] Likewise, as AI becomes integrated into healthcare systems, these systems may operate more efficiently, but often at the expense of decision-making on the part of both patients and

[9] See Betty Li Hou and Brian Patrick Green, "A Multi-Level Framework for the AI Alignment Problem," preprint, *arXiv.org*, January 10, 2023, doi.org/10.48550/arXiv.2301.03740.

[10] For a relevant discussion, see Matthew B. Crawford, *Why We Drive* (HarperCollins, 2020).

providers. In educational spaces, AI also has the capacity to restrict human agency at a systemic level. AI grading systems may replace the careful hand of a well-trained educator even while LLMs change the way students produce written work. Our political imaginations likewise stand to be shaped increasingly by algorithmic rather than human influences. The result is that AI has the capacity to undermine democracy itself.

Even as AI reshapes social, cultural, and environmental conditions, it will also increasingly exert an influence at the level of personal action and decision-making. Chapters 5 and 6 spent much space reflecting on these challenges, so we will not belabor the point here. Needless to say, as AI becomes increasingly powerful, these challenges to personal action and decision-making will only increase in scope and influence.

Finally, AI stands to challenge human agency at an even more basic level, in its integration with human biology. We are thinking here primarily of the rise of brain-computer interfaces (BCIs) and synthetic biological intelligences (SBIs). The former technology allows for the pairing of human brains with computers; the latter uses biological material in the construction of computational devices. Both are made possible in part by increasingly powerful AI, and several companies and labs are currently engaged in developing them. As these and similar technologies emerge, there is great capacity for them to be used for good in ways that complement or augment human agency. BCI technology, for example, can help provide assistive devices for patients with quadriplegia.

Yet the challenges to human agency posed by BCI and SBI technology are concerning indeed. Consider just one example: Wu et al. report development of a BCI that would detect neural activity associated with the urge to engage in binge-eating.[11] Having detected such an urge, the device could then be used to shut it down. Arguably, ridding patients of a desire to binge would allow them to overcome a tendency that poses a challenge

[11] Hemmings Wu, Sarah Adler, Dan Azagury, et al., "Brain-Responsive Neurostimulation for Loss of Control Eating: Early Feasibility Study," *Neurosurgery* 87, no. 6 (2020): 1277–1288, doi.org/10.1093/neuros/nyaa300.

to agency. But such a device would also raise important questions about the ownership of our own tendencies and agency. Would it be the *person* using the device who is overcoming the desire to binge? Or the device? The case also raises deeply concerning questions about how similar technology could be applied in other contexts. Should such a device be used to overcome fear? Or disgust? Or a nagging conscience? Clearly, in some applications, BCIs are poised to challenge human agency in a very basic way.[12]

Future Possibilities and Hope

AI has the capacity either to enable human freedom or to erode or undermine it. As we look to future possibilities, we therefore see tremendous challenges. We share the concerns that have been well articulated by many others, including the transhumanist thinkers introduced above. For example, we share transhumanist concerns about the capacity of future AI to undermine human flourishing and to pose serious existential threats.

We depart from transhumanists, however, in their ultimate aims and motivations. Transhumanists aim someday to eclipse limitations on humanity through technological means. As we discussed above, they thus exemplify what Pope Francis called "the technocratic paradigm" and what Pope Paul VI might have called a purely material hope. This brand of optimism about technological progress is grounded in an outsized estimation of both humanity and technology, and it is not a good fit with Catholic anthropology.

Yet Christians need not be completely pessimistic in regard to an AI-integrated future. In the face of questions about the future of AI and human agency, Christians are called away from the technocratic optimism of transhumanism and to something else instead: Christian hope. Hope, together with faith and charity, is foundational to Christian life and

[12] See similar discussion in The President's Council on Bioethics, *Beyond Therapy: Biotechnology and the Pursuit of Happiness* (President's Council on Bioethics, 2003).

morality. Hope will be particularly important in the face of the challenges we have covered here.

We could look to several places for an overview of what Christians mean by hope. Pope Francis articulates a particularly clear summary in *Spes Non Confundit*, the document decreeing 2025 a year of Jubilee:

> Everyone knows what it is to hope. In the heart of each person, hope dwells as the desire and expectation of good things to come, despite our not knowing what the future may bring. Even so, uncertainty about the future may at times give rise to conflicting feelings, ranging from confident trust to apprehensiveness, from serenity to anxiety, from firm conviction to hesitation and doubt. Often we come across people who are discouraged, pessimistic and cynical about the future, as if nothing could possibly bring them happiness. For all of us, may the Jubilee be an opportunity to be renewed in hope. God's word helps us find reasons for that hope.[13]

Christian hope, Francis goes on, is not a purely material hopefulness, but rather "is born of love and based on the love springing from the pierced heart of Jesus upon the cross."[14]

Faced with the challenges raised by AI, we are thus called to cultivate hopeful hearts, spirits neither of misplaced optimism nor of despondency and despair. We must live with the expectation of good things to come without slipping into naivete about the challenges we face. What does this kind of hope look like in action? That's where we turn now.

---

[13] Francis, "*Spes Non Confundit*: Bull of Indiction of the Ordinary Jubilee of the Year 2025," May 9, 2024, §1, vatican.va/content/francesco/en/bulls/documents/20240509_spes-non-confundit_bolla-giubileo2025.html.

[14] Francis, *Spes Non Confundit*, §3.

PART III
FOSTERING HUMAN AGENCY

While Part II criticized aspects of the current program for adopting AI, Catholic thought is not merely critical of technology. It has always encouraged the development of new technologies through which people can more effectively exercise their stewardship and support human flourishing.[1] In the modern era, Catholic social teaching (CST) has been the vehicle by which the Church has criticized new forms of technological society and suggested ways to positively appropriate these developments. Chapter 9 explores the resources of CST to address the challenges of AI for human agency, following Pope Leo XIV's call: "In our own day, the Church offers to everyone the treasury of her social teaching in response to another industrial revolution and to developments in the field of artificial intelligence that pose new challenges for the defence of human dignity, justice and labour."[2] Chapters 10 and 11 expand these basic principles into concrete suggestions. Through these chapters, we offer a picture of a flourishing society that uses AI in a way that affirms human agency.

[1] AI Research Group, *Encountering Artificial Intelligence,* 2–6; Green, "The Catholic Church and Technological Progress."

[2] Leo XIV, "Address of His Holiness Pope Leo XIV to the College of Cardinals," May 10, 2025, vatican.va/content/leo-xiv/en/speeches/2025/may/documents/20250510-collegio-cardinalizio.html.

CHAPTER 9

THE TREATMENT OF TECHNOLOGY IN CATHOLIC SOCIAL TEACHING

We ended Part II by considering the importance of Christian hope in plotting our technological future in a world with AI. In recent centuries, CST has given concrete guidance on how to live with hope in a rapidly changing world. From its first systematic articulation in *Rerum Novarum* (1891) to recent documents promulgated by Pope Francis, the social doctrine of the Church has offered a framework for how to reflect, how to judge, and when to act in such a way as to promote the dignity of all persons.[1] At the same time, this reflection must be done anew with each successive generation. As *Gaudium et Spes* summarizes this task:

> Though mankind is stricken with wonder at its own discoveries and its power, it often raises anxious questions about the current trend of the world, about the place and role of man in the universe, about the meaning of its individual and collective strivings, and about the ultimate destiny of reality and of humanity. . . . [T]he Church has always had the duty of scrutinizing the signs of the times and of interpreting them in the light of the Gospel. Thus, in language intelligible to each generation, she can respond to the perennial questions which men ask about this present life and the life to come, and about the relationship of the one to the other.

[1] Francis, *Laudato Si'*; Francis, *Fratelli Tutti*.

> We must therefore recognize and understand the world in which we live, its explanations, its longings, and its often dramatic characteristics.[2]

Since the Industrial Revolution, this need for reflection is increasingly tied to the development of novel technologies. While technologies have been present since the first humans, their development undertook an exponential increase in sophistication and proliferation beginning with the Industrial Revolution. From the first steam engine up to the most recent advances in AI, fundamental questions persist when scrutinizing a new technology: How should this new tool be used for the good of all? What does this new tool mean? How does this tool affect human dignity and agency?

At the same time, AI also poses fresh questions not yet examined by the tradition. In particular, as we explored in the earlier chapters of this book, AI's effects on human agency require new reflection on the relationship between individuals, technology, society, and God. Having presented many of these questions in Part II, in this chapter we will attempt to glean wisdom from the tradition of CST to provide hopeful guidance for an AI future.

As society increasingly looks to technology itself for the framework for a metaphysical "acceptable use policy," we find ourselves further and further immersed in a way of knowing the world that is predicated on efficiency, control, and optimization: an epistemology of technocracy. CST is a powerful counter to this, and it has argued since its inception that "all of the striving of men will be in vain if they leave out the Church."[3] It will be the intention of this section to take seriously the question "How have we and do we leave the Church *in*?" as technology continues to advance. This chapter traces a path from Pope Leo XIII to Pope Francis and the early days of Pope Leo XIV, taking into account key instances of technological progress within the framework of CST. The vision of

[2] Second Vatican Council, *Gaudium et Spes*, §3–4.

[3] Leo XIII, *Rerum Novarum*, §4.

technology that will emerge is one where the work of our hands represents true instances of "collaboration of man and woman with God in perfecting the visible creation,"[4] whether we are considering the steam engine or the computer.

Industrial Technology and Worker Rights

A significant focus of the encyclical *Rerum Novarum* was what this collaboration would look like as many elements of production industrialized. While the references to technology are implicit in this text, the beginning of a larger concern about defining the relationship between work, the person, and technology was clearly operative in the mind of Leo XIII. He writes:

> Hence, the employer is bound to see that the worker has time for his religious duties; that he be not exposed to corrupting influences and dangerous occasions. . . . Furthermore, the employer must never tax his work people beyond their strength. . . . His great and principal duty is to give everyone what is just.[5]

Leo makes it evident that the worker is *not* simply another piece of machinery in the ever-expanding industrial complex. For as industrialization expanded, as trade and production increased due to technological advancement and mechanized manufacturing, a reductive epistemology began to take hold. The person, in this way of seeing, was no longer different in kind from things and became, therefore, increasingly vulnerable to exploitation.

With this distinction between person and thing blurred, it became easier to reduce workers to the conditions of slavery, or close to it. The goals of unbridled capitalism, fueled by the creation of goods at increasing

[4] Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §2.
[5] Leo XIII, *Rerum Novarum*, §20.

speeds and decreasing costs, began to take precedence above the rights of the workers. The conditions that those in factories were subjected to, alongside the centralization of production and property, presented a landscape antithetical to what the worker was entitled to in order to flourish. Workers in this regime could not direct their own lives nor contribute to the common good; they increasingly lacked agency. Leo XIII responded to this growing concern by affirming both the right to property[6] and the right to a just wage and working conditions.[7] In so doing, clear boundaries for work, duties for the employer, limits on the state, and rights of the worker arise from within the framework of CST that was promulgated by *Rerum Novarum*. While this document spoke to a particular moment in human history, it has an enduring voice, especially for raising questions about the context in which technological developments occur and, in some cases, their unavoidable human cost.

Pope Pius XI, returning to *Rerum Novarum* forty years after its promulgation, spoke more specifically about a need to build social systems so that the resources afforded by industrial progress would remain accessible to all. The goods that technical achievement can afford ought "to be enough both to meet the demands of necessity and decent comfort and to advance people to that happier and fuller condition of life which, when it is wisely cared for, is not only no hindrance to virtue but helps it greatly."[8] Moreover, these same goods are not the sole responsibility of the state to direct. The "graduated order" described in *Quadragesimo Anno*, with its clear link to the CST principle of subsidiarity, was both a defense against the new technological powers of a potential totalitarian state and a vision for a new order that could foster harmony between increasingly disparate societal strata.[9] Such a vision, while having certain obvious

[6] Leo XIII, *Rerum Novarum*, §8, 9.
[7] Leo XIII, *Rerum Novarum*, §5, 20, 45–46.
[8] Pius XI, *Quadragesimo Anno*, §75.
[9] Pius XI, *Quadragesimo Anno*, §80, 83.

benefits for the state,[10] also effectively guards against the labor of the worker being reduced to a mere commodity. Labor as it *should* be viewed, as the fruit of workers' dignity, employs and enhances the fullness of their agencies by encouraging participation in decision-making, strengthening communities, and preventing state overreach.[11]

Global Technologies, Human Values

What emerged in the wake of the post-World War II technological expansion, while not entirely discontinuous from these questions of the rights of labor, was a new dimension of the social question. For Pope St. John XXIII, the work-person-technology relationship was framed by rapid technological developments, some of which imperiled the human family on a massive scale. As the mixed potential of emerging technologies began to become known, he asked how the fruits of human ingenuity could be used for the integral development of all persons.

In both *Mater et Magistra* and *Pacem in Terris*, John XXIII brought global technological advancement into sharp focus in the context of the nuclear age. It was the dawn of an age in which such technology was first used for great destruction, and then with a view toward promoting human flourishing and development. John XXIII comments:

> It pains us, therefore, to observe the complete indifference to the true hierarchy of values shown by so many people in the economically developed countries. Spiritual values are ignored, forgotten or denied, while the progress of science, technology and economics is pursued for its own sake, as though material well-being were the be-all and end-all of life. This attitude is contagious.[12]

[10] Cf. Pius XI, *Quadragesimo Anno*, §80, "Stronger social authority and effectiveness will be the happier and more prosperous the condition of the State."
[11] Pius XI, *Quadragesimo Anno*, §83.
[12] John XXIII, *Mater et Magistra*, §176.

Such an attitude presages the "technocratic paradigm" that would be a key element of Pope Francis's teaching almost fifty-five years later. What remains the alternative to this reductive anthropology? It is not, it must be said, a return to the Stone Age. Rather, what has been present and in a state of responsive equilibrium within the corpus of CST is a vision for what technology is meant *for*.

While the nuclear arms race precipitated this question for John XXIII, the validity of our present line of inquiry remains. We are, as the psalmist describes, created by God but charged with dominion of creation (8:5–6). CST is clear that "dominion" here is not a mere exercise of power without concern but a stewardship and caring for creation. With this framing, technology is a gift, one that is described as "part of the greatness of man that he can appreciate that order, and devise the means for harnessing those forces for his own benefit."[13] The heavens do tell the glory of God, and we have the potential to understand, to measure, and to cooperate with this glory for the good of all persons. For CST, this care for and celebration of the highest common good is the domain of technology and represents the *telos* of true progress.

These concepts, the proper domain of technology and the *telos* of "true" progress, remained operative throughout the Second Vatican Council and in the documents that followed. In particular, both the conciliar document *Gaudium et Spes* and Pope Paul VI's *Populorum Progressio* point to the reality that the advancement of technology is not value neutral. Indeed, the caution remains that what we value *most*, what imparts meaning to our lives, is not reducible to something that can be measured, consumed, and controlled. It is a very real danger that we may believe the lie that "in an era of scientific and technical triumphs such as ours man can well afford to rely on his own powers, and construct a very good civilization without God."[14] When God is left out, so is the true

[13] John XXIII, *Pacem in Terris*, April, 11 1963, §2, vatican.va/content/john-xxiii/en/encyclicals/documents/hf_j-xxiii_enc_11041963_pacem.html.
[14] Second Vatican Council, *Gaudium et Spes*, §57.

instantiation of our relationship to each other and to creation. The "dialog on man,"[15] the consideration of what is of value and meaning for all as connected to Someone and not a thing, is precisely Paul VI's concern:

> Man is truly human only if he is the master of his own actions and the judge of their worth, only if he is the architect of his own progress. He must act according to his God-given nature, freely accepting its potentials and its claims upon him.[16]

It is precisely because we can judge the worth of our actions, in a manner that is unique relative to all other forms of life, that the artifacts of what *persons* can make can be truly categorized as progress. So far, the vision fashioned by CST for how we are to see, judge, and act where the work of our hands is concerned points to a theology of progress that includes the promotion of the integral development of all persons, equality of access, the need for discernment, and moral evaluation and formation.

Remaining Human in a Digital Age

How then, Pope St. John Paul II would go on to ask, should we continue to view technological advancement in a future that begins to include attempts to delegate uniquely human capacities to machines and to replicate, synthetically, the very essence of our bodily reality? While he astutely drew upon the teaching of *Rerum Novarum* in *Centesimus Annus*, in *Laborem Exercens* John Paul II returned to the fact that the dignity of work was irreducible to the complexity of a given task. How persons are to live out their vocation *qua persons*, that is to rule over creation, is inseparable from action and creation in labor.[17] For it is "only man [who] is capable of work and . . . thus work bears a particular mark of man and of humanity, the mark of a person operating within a community of

[15] Paul VI, *Populorum Progressio*, §73.
[16] Paul VI, *Populorum Progressio*, §34.
[17] Genesis 1:26.

persons."[18] The work that is asked of a person, how this work relates to dignity, and what tool is used to accomplish that work, are accordingly incredibly important considerations. While the advance of technology has slowly catalyzed a reductive epistemology rooted in technology, CST defends the position that the work a person does is inseparable from the dignity that person has and the flourishing people ought to realize.

This resistance became especially important as, during the pontificate of John Paul II, the Church discerned its response to the next technological leap: biotechnology. Beginning with the distinction between a "culture of life" and a "culture of death," the operative question again was about the relationship of person-work-technology and what the true *telos* of technology should be. As we attempt to rightly order this relationship, John Paul II pointed out a building tension between these two "cultures."

> The heart of the tragedy being experienced by modern man: the eclipse of the sense of God and of man, typical of a social and cultural climate dominated by secularism, which, with its ubiquitous tentacles, succeeds at times in putting Christian communities themselves to the test. Those who allow themselves to be influenced by this climate easily fall into a sad vicious circle: when the sense of God is lost, there is also a tendency to lose the sense of man, of his dignity and his life; in turn, the systematic violation of the moral law, especially in the serious matter of respect for human life and its dignity, produces a kind of progressive darkening of the capacity to discern God's living and saving presence.[19]

While in this case John Paul II was speaking predominantly to the inception and rapid proliferation of biotechnology, we can continue to make the claim that the "two cultures" distinction re-presented a common theme in CST. Namely, CST explores what confers upon all persons their dignity as created in the *imago Dei*: relationship. Being created in relationship to our Creator and, therefore, in relationship to each other

[18] John Paul II, *Laborem Exercens*, Introduction.
[19] John Paul II, *Evangelium Vitae*, §21.

and to God's creation, bestows upon all persons a solidarity, a commitment to the good that we should all value and share in common. But if we do not share in a common commitment to this good, how can we maintain an acknowledgment of the *imago Dei* we each possess?

This is what Benedict XVI was thinking about in *Caritas in Veritate*. At the very heart of the Church's social doctrine, he wrote, is charity. This chief virtue does not involve acquiescing in a relativistic sense to whatever another affirms to avoid offense. There is, rather, a profound link between truth objectively understood and charity.

> Truth is the light that gives meaning and value to charity. That light is both the light of reason and the light of faith, through which the intellect attains to the natural and supernatural truth of charity: it grasps its meaning as gift, acceptance, and communion. Without truth, charity degenerates into sentimentality.[20]

It cannot be stressed enough that without both truth and charity, technology becomes a tool for the powerful and a weapon against the poor and marginalized. When dissociated from truth and charity, technology is the basis for a mindset that is bent on efficiency at the expense of relationality. With a right association to truth and charity, technology presents to the human family the opportunity to exercise stewardship over creation with increasing sophistication, reduce the risks and time of certain types of labor, and improve working conditions; it is, in two words, fascinating and freeing. But a good response to technology must remain within the moral law and thus in accord with truth and charity. We must move beyond the illusion of total autonomy afforded by a misapplied vision of technology and toward a reality "that He has entrusted to humanity, and it must serve to reinforce the covenant between human

[20] Benedict XVI, *Caritas in Veritate*, §3.

beings and the environment, a covenant that should mirror God's creative love."[21]

The Technocratic Paradigm

Under the pontificate of Francis, CST once again looked afresh on the questions of the day. Following his predecessors, Francis reflected on the outsized role that technology has in our lives and worried that the values espoused by CST are not the same as those motivating the design and usage of most modern technologies. In fact, Francis carried the tradition's critical analysis of modern technology further than his predecessors in two key ways. First, Francis was the first to explicitly address AI as a primary subject for reading the "signs of the times" starting in the 2010s.[22] Second, his discussions of AI in public addresses must be contextualized within Francis's more extensive contribution to the present conversation, namely his delineation of the technocratic paradigm and the ways it shapes our understanding of and relationship with technology today.

When Francis issued *Laudato Si'* in 2015, it was praised around the world as the first encyclical to address the global climate crisis, but within his argument about the moral duty to care for our earth, Francis also articulated the best-developed ethic of technology in CST to date. At its foundation is an observation that the world is currently operating with a particular set of assumptions about how our socioeconomic structures should work and a particular understanding about the relationship among

[21] Benedict XVI, *Caritas in Veritate*, §69.

[22] For a brief history of the Vatican's internal work on AI, see Brian Patrick Green, "The Vatican and Artificial Intelligence: An Interview with Bishop Paul Tighe," *Journal of Moral Theology* 11, Special Issue 1 (2022): 212–231, jmt.scholasticahq.com/article/34131-the-vatican-and-artificial-intelligence-an-interview-with-bishop-paul-tighe. For Pope Francis's first official allocution mentioning AI, see Francis, "Message of His Holiness Pope Francis to the Executive Chairman of the 'World Economic Forum' on the Occasion of the Annual Gathering in Davos-Klosters," January 23–26, 2018, vatican.va/content/francesco/en/messages/pont-messages/2018/documents/papa-francesco_20180112_messaggio-davos2018.html. It is worth noting that Pope Leo XIV has indicated that he will pick up this mantle and that AI is set to be a central topic for him in the coming years: Leo XIV, "Address to the College of Cardinals," May 10, 2025.

the natural world, humanity, and technology. As we have shown above, elements of this set of assumptions and understandings have been observed by practicing CST's core tenet of "reading the signs of the times" since *Rerum Novarum.* Francis built on these observations to show how these ideas form a *paradigm* that supports and preserves this problematic framework.[23] Using this paradigm, we have shaped an entire understanding of how human life, society, and creation as a whole both do and should work. This "epistemological paradigm . . . shapes the lives of individuals and the workings of society"[24] and, importantly for us, it defines "the way that humanity has taken up technology and its development."[25]

However, the nature of a paradigm is that it is essentially invisible to those who operate within it. Its very ubiquity makes it difficult to distinguish between facts and the assumptions of the paradigm because all observations are interpreted by means of it. These assumptions are often easier to see after a paradigm has been seriously challenged or overturned. For example, at one point we believed that the entire universe revolved around the earth. The assumptions built into the geocentric paradigm were taken as fact, and to the everyday observer, it was impossible to conceive that the sun was not a bright sphere that moved across the sky but rather a body around which the earth orbits and whose relative motion in our sky is actually a function of the *earth* revolving. Now that this paradigm has been overturned, it is much easier to see the construction of the understanding that allowed people to accept a very different model in which our planet occupies a minor place circling a medium-sized star.

Francis claimed that the same power of paradigmatic thinking is operating upon our culture today: "The idea of promoting a different cultural paradigm and employing technology as a mere instrument is nowadays inconceivable. The technological paradigm has become so

[23] Francis, *Laudato Si'*, §106.
[24] Francis, *Laudato Si'*, §107.
[25] Francis, *Laudato Si'*, §106.

dominant that it would be difficult to do without its resources and even more difficult to utilize them without being dominated by their internal logic."[26] The technocratic paradigm sets our values and shapes our decisions. It determines everything—how engineers design, how marketers market, and how users use products. It even shapes our choices for the discarding and disposal of that same technology. Francis names the technocratic paradigm as the source of many of the injustices that plague our communities in the twenty-first century.[27]

In naming the assumptions upon which our social life is built as a "cultural *paradigm*," Francis made a very specific claim about what prevents us from dealing with the threats of climate change (in *Laudato Si'*) or the AI revolution (in later writings). We need more than just a realignment of our values. Rather, we need to see through the false assumptions upon which our entire social world is constructed. Though he never uses the term, what Francis is calling for is a full *paradigm shift*. Of course, the first step toward a paradigm shift is recognizing that we are operating within a paradigm, and that the assumptions of this paradigm are subject to review. Throughout his writings, Francis included countless observations about the technocratic paradigm.[28] Time and space do not allow for a full accounting here, but it is worth examining how Francis described the assumptions the technocratic paradigm makes about human agency.

Even before he identified the technocratic paradigm, Francis sowed the seeds for this development by demonstrating the interconnection between personal choices and structural forces. This can be seen most clearly in his first apostolic exhortation, *Evangelii Gaudium*:

---

[26] Francis, *Laudato Si'*, §108.

[27] Cf. Lebacqz and Gaudet, *Eight Theories of Justice*, 113–117.

[28] While he often used the term "technocratic paradigm" directly, Francis also substituted other terms such as "throwaway culture" or "an economy that kills" to emphasize certain aspects of the paradigm.

> Just as the commandment "Thou shalt not kill" sets a clear limit in order to safeguard the value of human life, today we also have to say "thou shalt not" to an economy of exclusion and inequality. Such an economy kills. How can it be that it is not a news item when an elderly homeless person dies of exposure, but it is news when the stock market loses two points? This is a case of exclusion. Can we continue to stand by when food is thrown away while people are starving? This is a case of inequality.[29]

Francis reframed the command "*Thou* shalt not kill," typically understood in terms of an individual's actions (that is, *I* should not pull the trigger), as an entire "economy that kills" (that is, we have *all* pulled the trigger when a person dies of exposure). At the same time, Francis did not resign the problem to society as a whole; we are *each* responsible for the social conditions that lead to homelessness and hunger. He identified a vicious cycle between the impersonal social structures of our present world and a lack of empathy we have for the suffering other.

In *Laudato Si'*, Francis connected this relationship between personal choices and social sin to the outsized role that technology today has in determining the nature of the lives we live and the societies in which we operate: "We have to accept that technological products are not neutral, for they create a framework which ends up conditioning lifestyles and shaping social possibilities along the lines dictated by the interests of certain powerful groups. Decisions which may seem purely instrumental are in reality decisions about the kind of society we want to build."[30] Francis was particularly distressed by the ways in which our interactions with technology have eclipsed our humanity. In the technocratic paradigm, technology has become an end unto itself: "Technology tends to absorb everything into its ironclad logic, and those who are surrounded with technology know full well that it moves forward in the final analysis

[29] Francis, "*Evangelii Gaudium*: Apostolic Exhortation on the Proclamation of the Gospel in Today's World," 2013, §53, vatican.va/content/francesco/en/apost_exhortations/documents/papa-francesco_esortazione-ap_20131124_evangelii-gaudium.html.
[30] Francis, *Laudato Si'*, §107.

neither for profit nor for the well-being of the human race.'"[31] In contrast, he argued that technology should never be valued as an end unto itself. Technology should be a tool: instrumentally valued for its usefulness in achieving other, intrinsically valuable, ends.

In our pursuit of technology, some people also seek "progress" as if it were an intrinsic good and without asking questions about the benefits or costs of our pursuits. They might desire new products not because they are useful, but for their own sake. They might want the latest smartphone model even though the marginal difference from an existing phone is imperceptible. They might replace appliances with new models that are "smart" even if they do not need and rarely use these features in real life. And most recently, every product seems to tout the integration of AI, regardless of whether that integration improves the product's utility.

As we take those things which should be valued instrumentally and give them intrinsic value, we also do the converse. It is a central teaching of CST that human life and dignity are to be treated as intrinsically valued and inviolable. Today, as in the time of *Rerum Novarum*, "human beings are themselves considered consumer goods to be used and discarded. We have created a 'throw-away' culture which is now spreading."[32] We treat other human beings as solely "human resources" and creation as solely "natural resources," all to be consumed as necessary for the sake of technology, progress, disruption, and profit themselves.

Moreover, as the role of technology in our lives increases, our sense of our own humanity decreases: "Our immense technological development has not been accompanied by a development in human responsibility, values and conscience. Each age tends to have only a meager awareness of its own limitations. It is possible that we do not grasp the gravity of the challenges now before us."[33] We have power, but, in Francis's view, we have not learned to use that power well.

---

[31] Francis, *Laudato Si'*, §108.

[32] Francis, *Evangelii Gaudium*, §53. See also Francis, *Laudato Si'*, §20–22.

[33] Francis, *Laudato Si'*, §105, quoting Romano Guardini, *The End of the Modern World*.

Catholic Technological Teaching

Despite its continual cautions about the misuse of technology in every era since *Rerum Novarum*, CST is not anti-technology. In fact, it could be argued that one of the great themes of the tradition has been providing guidance on how to best utilize technological innovation.[34] In Francis's words: "The Church never ceases to encourage the progress of science and technology placed at the service of the dignity of the person, for an 'integral and integrating' human development."[35] To accomplish this, however, it may be helpful to excerpt and organize a "Catholic technology teaching" that can help guide us toward better development and use of technology today.

First, all technology needs to be developed with an aim to respect **human dignity** and support human flourishing. We should be cautious about any technology that does not ultimately serve these ends. Human flourishing includes the care for human wellbeing (physical, psychological, and spiritual), the development of lives of purpose and meaning, and the encounter of beauty and happiness that inspire genuine wonder and joy.

Second, **human relationships** are a necessary part of human life. Technologies that aim to replace genuine human contact with human-technological interactions should be pursued with extreme caution. There may well be use cases in which replacing humans with machines leads, ultimately, to positive outcomes (e.g., robot drivers or surgeons may surpass their human counterparts in the ability to keep humans healthy and safe), but for many applications, the goal of "efficiency" has a tendency to eclipse all other concerns, including the genuine need for human interaction and the social need to connect and empathize with one another. Francis cautioned us to use our technological agency to seek a "culture of encounter," rather than an isolated existence.

---

[34] Green, "The Catholic Church and Technological Progress."

[35] Francis, "Address of His Holiness Pope Francis to the Members of the Pontifical Academy for Life," February 20, 2023, vatican.va/content/francesco/en/speeches/2023/february/documents/20230220-pav.html.

Next, we need to acknowledge, with Francis, the existence of a **technocratic paradigm** that frames our cultural priorities and definitions of technological success. To resist this damaging paradigm means breaking away from the notion that technology "advances" in a solely positive direction. This is not a new realization. The cautions levied against nuclear weapons in *Pacem in Terris* and *The Challenge of Peace* lament that the development of such weapons left the world, in fact, far less secure.[36] *Fratelli Tutti* similarly weeps at the lack of attention paid to the psychological and social costs of our digital technologies. But today, as the pace of technological innovation seems to move faster than human comprehension, there is an even greater need to make space for caution and care when it comes to the harmful effects of each new technology.

Of course, CST has, from its outset in the Industrial Revolution, been a teaching on technology. Thus, CST's central concepts and principles—**solidarity, the common good**, **subsidiarity**—remain vital for any formal Catholic technology teaching as well. One cornerstone of CST, the **preferential option for the poor and vulnerable**, must be upheld in considerations of technology, especially in light of the growing inequality between the technocratic elites and the consumers who do not create the technologies they purchase. Ensuring that technology is giving priority to the most vulnerable, not the technologists' profits, is crucial.

Finally, we need to remember that our capacity to create technology emerged hand in hand with our responsibility to care for creation. With technology consuming exponentially increasing amounts of energy and clean water, we must heed our **care for our common earth** as a necessary precondition for building or expanding any technology.

Over the course of its history, CST has suggested many tools for protecting human dignity and transforming our worldview. Some of these are the very personal exercises of the spiritual life: fostering individual

[36] John XXIII, *Pacem in Terris*; The National Conference of Catholic Bishops, *The Challenge of Peace: God's Promise and Our Response* (United States Catholic Conference of Catholic Bishops, 1983).

conversion, spiritual devotion, and acts of solidarity. CST also suggests ways for societal intervention. Some of these steps involve legal and regulatory enactments by governments, such as the basic protections of workers for which Leo XIII argued. The documents of CST also suggested new forms of social life beyond the government and market that could address these problems: labor unions, cooperatives, business marked by gratuitousness, and different forms of civil society organizations. The coming chapters explore ways to address the challenge of AI for human agency that are inspired by CST. Chapter 10 examines legal and regulatory enactments that might be necessary to defend human agency by encouraging transparency, attention, and accountability. Chapter 11 explores what new forms of organization or technological design might better protect human dignity. These suggestions present a vision for AI that serves human agency.

Chapter 10

Constraining the Negative Effects of AI on Human Agency

As the last chapter discussed, CST provides a positive vision for society's engagement with new discoveries and human creations, but its documents also have suggested concrete moral guidance to limit the negative effects of technologies and adverse social arrangements. CST recognizes a role for governments in restricting harmful actions and regulating deleterious social practices, while also embracing the need for individuals and collectives beyond government to respond to these same problems. This chapter describes four areas in which current practices of AI development and use need to be constrained. First, governments and companies must take steps to limit the manipulative nudging that is central to surveillance capitalism. Second, individuals and communities must limit the effects of the attention economy on their ability to focus on action and the world around them. Third, these steps would be aided by increased transparency surrounding when AI is used, how it is used, and its effects on the user and those subject to AI-managed processes. Finally, increased transparency should be matched by a right of appeal when an AI application has led to a deleterious outcome for a person. A commitment to transparency should also emphasize the need for human judgment, in the first place, in important spheres of life, like medicine or criminal justice, where mistakes can lead to catastrophic outcomes. While there are complications in defining the appropriate role of AI in each of these areas, constraining AI applications in these ways would help to protect responsible human agency.

Limiting Nudging

A first set of necessary restrictions concern technologies that attempt to manipulate people using psychological techniques that take advantage of vast amounts of individualized data collected by modern computers. As chapter 5 discussed, programmers can influence human action through targeted nudges or capture it wholesale through addictive technologies. In the latter case, addiction can partially or more completely undermine agency, severely reducing the human ability to freely act in the face of an urge.[1] And although nudging can be used in a way that is respectful of human agency, when it is used in a highly individualized, AI-dependent way, it can also undermine agency by intentionally short-circuiting the deliberative process. It seeks to hide reasoning about action from people. The danger of these addictive and manipulative techniques is accentuated because they are largely hidden, making it difficult for us as individuals to protect ourselves from the assaults of the large corporations and institutions that use them. Their hidden nature and power make it imperative for legal and regulatory institutions to confront these practices, even if companies should also seek to avoid these techniques.

Yet, currently proposed ethical and legal remedies fail to address this manipulative infrastructure. Few of the many sets of "principles of AI ethics" reference nudging or addictive technologies. Even the strongest regulations tend to ignore aspects of addictive and manipulative technologies. For example, the EU "AI Act" only would apply remedies for manipulation in particularly high-risk domains, not those of advertising or video games, though the former are famously manipulative and the latter are frequently highly addictive.[2] Though they address some abuses, many sets of regulations and ethical principles would leave in place an ecosystem in which human attention is absorbed by devices. While we

[1] This assault on human free will is indicated even by the etymological origins of the word "addiction," which comes from a Latin legal term by which a person takes possession of a slave, making the slave "spoken for."

[2] *The Act Texts | EU Artificial Intelligence Act*, 2024, artificialintelligenceact.eu/the-act/.

call for broader regulatory efforts, in this section we will primarily discuss two aspects of these technologies that raise particular concern: the need to regulate advertising and the need to protect attention.

Advertising

Part of the problem with current government interventions is that regulatory frameworks were not built to address this completely new system of advertising. In the past, marketing was in many cases highly personalized, in the sense of a salesman guiding the customer to a particular product. Alternatively, it was highly systematized at the level of the population, with examples including mass advertisements or putting commonly used items like bread at the back of the store so people would have to walk past other products. This separation between personalization and systematization created important limits on the marketer. With AI, however, personalized selling can be systematized.[3] A customer is invited to an addictive game at the precise time that they are most likely to accept it (Wednesday afternoon for Steve or Friday morning for Brenda). While regulators have confronted some earlier forms of abusive marketing, they have not yet addressed this new, powerful combined model.

Marketing, and specifically advertising, is perhaps the feature of capitalism that is subject to the most intense criticism. This criticism is justifiable when a company's tactics resemble manipulation rather than persuasion, where manipulation must be regarded as an assault on agency. Convincing a person to do something is not itself the problem. Persuasion is obviously good and essential for any human life, so if there is a way for a company to offer a consumer compelling reasons to engage in a transaction, and that consumer can consent without duress, a free society should respect that consumer's decision. The problem is that in practice, it is often extremely difficult to distinguish between persuasion and manipulation. They are not two distinct concepts but rather exist on a

[3] Ryan Calo, "Digital Market Manipulation," *George Washington Law Review* 82, no. 4 (2014): 995–1051.

continuum: There are relatively few purely manipulative or purely persuasive interactions. An unfortunate implication is that for any transaction, a company can almost always say, "See, we persuaded the consumer to purchase this product," and a consumer advocate can almost always say, "No, you manipulated the consumer into purchasing that product."

As noted, this is an old dynamic, and it has existed as long as there have been markets,[4] but this problem now has new dimensions because of the capabilities of digital technology. Salespeople have always attempted to manipulate individuals based on their identified features, and corporations have always attempted to manipulate the masses through psychological techniques, but in the pre-digital era, the personal/masses distinction was a meaningful boundary. Algorithms have essentially erased that boundary.

A first step is to examine what makes an attempt to convince customers a manipulative versus a persuasive operation. Any attempt to convince will be based on the company's understanding of the needs and wants of its potential consumers. But there are two very different ways to complete this transaction. Consider first a company that persuades a consumer that the company could sell them goods or services that met their needs. For example, if I were headed out to hike in the desert, I would no doubt be aware of needing water. But I might be quite happy for someone to point out other things I could benefit from, such as good footwear, sunscreen, a hat, and so on. A company that advertised these things should be described as a company helping me to discover my own needs and wants.[5]

But a company can also use advertising to create the needs and wants in the first place. For example, perhaps John is content with the older car he has but becomes discontented because of a well-designed advertisement

[4] Outside of markets, the problem has long bedeviled questions about political persuasion enacted through rhetoric, as in Plato's criticisms of the sophists in dialogues like *Gorgias* or *Phaedrus*.

[5] John Mackey and Raj Sisodia, *Conscious Capitalism: Liberating the Heroic Spirit of Business* (Harvard Business Review Press, 2014).

that associates having a different car with gaining social influence or connections. The advertisement is not helping him discover anything; it is creating a desire in John that does not need to be there. An advertisement that helps a person discover needs and wants should be classified as persuasive; an advertisement that *creates* needs and wants should be regarded as manipulative. The former seeks the good of the customer while also benefiting the company, whereas the latter practice seeks only the good of the company. Again, in practice, this difference can be difficult to see, but the distinction is ethically critical vis-a-vis human agency.

Questions to Ask

The task of distinguishing manipulation and persuasion is in practice quite difficult because, while we can construct their respective features theoretically, in real world examples things often depend in large part on the way they are described. Still, though, there are questions that we can ask, considerations we can reflect upon, that can clarify whether something is persuasive or manipulative, or whether a technology makes an inappropriate demand for attention. In these following paragraphs, we offer a list of questions that may serve as a starting point for constructing meaningful distinctions. In each question, you should think of the word "persuade" in quotes, because the very question is whether "persuasion" or "manipulation" is a more appropriate description. These questions can serve as initial steps for designers to evaluate their advertising technologies.

1. Is the subject aware that there is an attempt to persuade? If the subject is unaware of this attempt, then it is more likely that the interaction is manipulative.

2. Is the persuasion directed at the conscious/deliberative part of the subject? There are generally two methods of decision-making: a quick, intuitive way, and a slow, deliberative way.[6] Humans are most

[6] Daniel Kahneman, *Thinking, Fast and Slow*.

likely to reflect on whether what is offered is actually good for them if they decide in a deliberative way. If an attempt to convince someone of something appeals to the mental shortcuts of quick decision-making tendencies, then is it likely manipulative.

3. Is the product or service that the persuading company wants to offer a product that will be beneficial for the subject? Though this question may be difficult to answer because people will argue endlessly about what is good for health, wellbeing, or virtue (even cigarette manufacturers will have a story), it is always worth asking.

4. Does the persuasion depend on timing? If the advertisement is delivered to be most effective at a particular time, then it is almost certainly attempting to avoid our rationality and hence is manipulative.

5. Does the persuasion depend on the emotional state of the subject? If an appeal is more effective if a person is angry, afraid, or sad, and the advertising effort strives to create the emotional conditions necessary for the good or service to be more appealing to the subject, then it is almost certainly manipulation.

6. And finally: Is the persuasion making such consistent demands on attention that the person is unable to focus on other things? Such a demand on attention overwhelms the deliberative faculty more globally, not just affecting one consumer choice but the whole network of decisions a person can make.

Attention

Many advertising and AI technologies more generally trade on attention, which itself is a significant threat to human agency because it is perhaps the most fundamental basis for agency. How we exercise our agency is

always first dependent on what we turn our attention to. Robert Spaemann, building on an older tradition of philosophy, argues that human freedom depends on our attention. As he writes, "The problem of the freedom of the will . . . is apparently that of focusing our attention."[7] Human action is always for some good that appears in the surrounding world. Yet there is always more than one good for which a person can act in the environment. What people "can do is direct their attention on an object, fill their thoughts and imagination with it, regardless of life's necessities, and concentrate more protractedly upon it than they would have done without a definite resolve."[8] The problem of sin is that people act for a lesser good, like money or ambition, in place of a higher good, such as God or our neighbor's needs. As Augustine painted our situation, our loves are disordered; we prioritize lower over higher goods.[9] When faced with temptation, freedom for the good requires turning to the higher good. Virtuous action is therefore not so much the crushing of bad action or desire as the redirection of attention toward good action. Attention is thus the heart of human agency.

It is this fact that makes the attention economy so dangerous.[10] When overwhelmed by images, people lose the ability or the strength to control their attention. Moreover, others use the incredible power of AI, combined with psychological techniques, to bend attention to their will. In so doing, they undermine human agency more broadly. It is thus imperative to protect attention.

In many ways, developing attention is a personal project—and as with all personal projects, it also must rely on the help of God's grace. Even

---

[7] Robert Spaemann, *Persons: The Difference Between 'Someone' and 'Something,'* trans. Oliver O'Donovan (Oxford University Press, 2006), 219.

[8] Spaemann, *Persons*, 219. Here he draws on William James's understanding of the will.

[9] E.g., Augustine, *Confessions*, trans. F. J. Sheed (Sheed & Ward, 1947), 2.5.10–11; Augustine, *The City of God Against the Pagans*, ed. R. W. Dyson (Cambridge University Press, 1998), 15.22.

[10] Matthew B. Crawford, *The World Beyond Your Head: On Becoming an Individual in an Age of Distraction* (Farrar, Straus and Giroux, 2016), 3–30.

without AI or advanced technology there will always be some other good in the world to distract us, some other pastime in which to lose ourselves. We must make an effort to direct attention. No other service or technological program can give us back to ourselves. However, technology *can* play an important supporting role in a person's quest for attention: Internet blockers can filter out distractions; apps can make us aware of new prayers; while still too early to fully ethically evaluate, neuro-technologies are seeking to indicate when the user becomes distracted. Though not the complete answer, technology development in support of attention should be encouraged, though not without scrutiny. Designers can even engage a broader consideration of the way that any technology or program will affect attention as a part of a technology's design.

Even though individual efforts are necessary, today's attention economy is so broad and powerful that it is a social problem. It is an environment that makes it hard to develop virtuous agency and thus undermines the common good. In this situation, laws and regulations become necessary. These are most needed for the young, whose powers of attention are still developing. Recent years have seen movement on this front, with some proposals for social media bans for children under thirteen or bans on cellphones in schools. More explorations along these lines are necessary. Regulatory actions could also target specific techniques, such as autoplay of the next video, notifications that distract users, or manipulative nudges. Even if not all techniques of attention demand legal regulation, they can still be subject to industry rules for best practices and internal regulations. These kinds of formal controls are necessary because of the social nature of attention.

Transparency

Greater transparency surrounding the use of AI would help prevent the manipulation of choice and attention, but it would have broader effects as well. It might seem evident that awareness is integral to any reflection on human agency because it is a precondition for conscious action and for

protecting moral agents from abuses and manipulations that deprive us of our agency and, in so doing, undermine our dignity. Awareness might mean knowing what is happening and the implications for individuals, society, and the planet, as well as knowing what is concealed or omitted. In the case of AI, this awareness implies that people and institutions should know when they are subject to AI, when they are interacting with AI systems, and how AI systems operate, at least at a level that protects individual and institutional agency. Transparency should make such a multilayered awareness possible.[11]

Discussions of transparency often presuppose a degree of technological literacy that, on the one hand, allows for identification of the presence of AI and its uses and, on the other hand, articulates ways that AI applications are considered acceptable or even needed. An example of this approach is the 2024 "AI Act" of the European Commission, according to which "transparency means that AI systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights."[12] In light of this approach, "transparency is required for all AI relevant systems."[13] The first ethical goal to be pursued is "to establish a reliable and transparent regime for AI technologies, which is instrumental in enabling market operators as well as individuals to understand the AI systems' design and use."[14] The second

[11] "The principle of *transparency* points to the danger that users frequently do not know how an AI system determines its recommendations (or even that an automated system is making a decision). . . . Contemporary AI often uses deep neural networks with hidden layers of processing. This can make it difficult to understand the reasoning behind a decision, which in turn makes it harder to root out errors in AI decisions. Thus, many approaches to AI ethics that focus on principles have suggested that a certain amount of transparency is necessary to build into AI systems." See AI Research Group, *Encountering Artificial Intelligence*, 26.

[12] European Union, "AI Act," Recital 27, euaiact.com/key-issue/5.

[13] European Union, "AI Act," Recital 27.

[14] European Union, "AI Act," Article 3(63), digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers.

goal is to promote "the responsible development and use of AI, strengthening the accountability of relevant market actors for their AI operations."[15]

Furthermore, a case-based assessment is needed because "the transparency requirements . . . vary depending on the system's risk level."[16] Regulations are needed and "the applicable rules must be identified on a case-by-case basis after a meticulous examination and taking the special circumstances into account."[17] This said,

> The EU AI Act does not clearly address the actual level of transparency and understandability that will be required for AI systems. In other words, it is not yet clear in practice how and to what extent it will be sufficient to comply with the EU AI Act rules on transparency. The practical tools, templates, and particularly codes of practice are to be developed by the AI Office. However, the true meaning and interpretation of transparency will be shaped through the implementation of the Act in time.[18]

To summarize, for the "AI Act," transparency is necessary and should be defined on a case-by-case basis, but which level of transparency is demanded, and which type of compliance is required, are still uncertain and will be developed in the future. Transparency here, it seems, is an aspiration that longs to be actualized.

The Christian ethical tradition shares some of these concerns about transparency but adds a greater awareness of the complex elements that contribute to agency, such as the role of ignorance. External factors influence our decision-making process and then our actions, especially by affecting our access to knowledge. The Catholic moral tradition breaks down the term "ignorance" in ways that can illuminate the importance of transparency for action. For example, *antecedent ignorance* describes when

[15] European Union, "AI Act," Article 3(63).
[16] European Union, "AI Act," Recital 27.
[17] European Union, "AI Act," Recital 27.
[18] European Union, "AI Act," Recital 27.

a lack of knowledge precedes decision-making and influences it, as well as the moral assessment of the consequent decisions and actions. An agent is not culpable for antecedent ignorance, but a social structure can lead agents to harmful actions if it withholds information from an agent.[19] To reclaim transparency regarding AI to promote and protect agency aims at limiting and, as much as possible, at avoiding antecedent ignorance.

A social structure should also allow us to seek the information we need to avoid *consequent ignorance*, which describes what is happening when we purposefully avoid, whether out of laziness, neglect, or bad intentions, gathering information that we have a responsibility to seek in order to fulfill obligations or act safely.[20] It is easy for a complex society to make excuses for not providing proper information, for example by claiming that it requires too much effort or too much literacy. Societies should avoid making it overly burdensome to acquire the information we need to make an informed decision. Structures should support agency in such a way that allows the community to receive the information required to determine how needs can be met and to what extent a person can ethically cooperate with communal action and to identify vulnerable populations that need special protections.

Transparency also has implications for fairness, an emphasis of existing Catholic thought on AI. Hence, "Most generally agree that AI should be built with fairness in mind, but complex debates have been waged over what counts as 'fair.'"[21] Such complex discernment processes cannot be avoided, and though they may be challenging to conduct, they are certainly worthwhile. In pursuing fairness, people should give attention first to the technology, i.e., to the algorithms which are produced and used, with their specificity and limitations; second, to the role that regulations have; and third, to the moral agents. The document *Antiqua et Nova* states, "Insofar as AI can assist humans in making decisions, the algorithms that govern it

[19] Aquinas, *ST* I-II, q. 6, a. 8.
[20] Aquinas, *ST* I-II, q. 6, a. 8.
[21] AI Research Group, *Encountering Artificial Intelligence,* 29.

should be trustworthy, secure, robust enough to handle inconsistencies, and transparent in their operation to mitigate biases and unintended side effects."[22] Moreover, "regulatory frameworks should ensure that all legal entities remain accountable for the use of AI and all its consequences, with appropriate safeguards for transparency, privacy, and accountability."[23] Finally, "Those using AI should be careful not to become overly dependent on it for their decision-making, a trend that increases contemporary society's already high reliance on technology."[24] Hence, for Catholic social thought, discerning which type of transparency is fair implies avoiding deception,[25] manipulation, misrepresentations,[26] exploitation,[27] and respecting people's dignity.

These concerns extend not only to how individuals engage in these technological structures but also to how technological systems utilize the data they extract from human users. Decision-making workflows in many spheres have already been subtly or radically altered by the integration of AI, distancing humans from engagement with pertinent information. Data is also collected in discreet ways through our everyday interactions with technology and is analyzed for decisions that affect product promotion, marketing, public health, relationships, and more. This is all despite the fact that we may never be fully aware of the downstream uses

---

[22] Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §46. See also Francis, "Address to Participants in the 'Minerva Dialogues,'" March 27, 2023, vatican.va/content/francesco/en/speeches/2023/march/documents/20230327-minerva-dialogues.html.

[23] Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §46; see also §93. See also Francis, "Address to Participants in the 'Minerva Dialogues.'"

[24] Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §46. See also Francis, "Address to Participants in the 'Minerva Dialogues.'"

[25] See Dicastery for the Doctrine of the Faith, *Dignitas Infinita*, §1; see also Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §62.

[26] See Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §84.

[27] See Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, §93.

of our data and the implication of our personal information in wider economic schemes. To respect the dignity proper to human freedom and prioritize human prudence over the output of AI, it is important, as the next section discusses, for societies to preserve an appeal process for decisions from an AI algorithm and a means for understanding the uses to which AI is applied. Maintaining rights to appeal and access to information that was used for making decisions safeguards a person's vulnerability to coercion and exploitation, protects safety, respects human agency, and promotes human participation in a community.

Maintaining the Human Role and the Right of Appeal

Algorithms can collate large data sets to help communities execute thoroughly informed plans of action. Yet, decisions by algorithms may lead to deleterious consequences when they circumvent the wisdom of the human person gained from lived experience or simply mental flexibility.[28] For example, algorithms utilized in the clinical setting can indicate various health risks for particular patients, including when a patient is at risk of dying within the year.[29] This scenario creates a dilemma for the physician, who must decide whether or not it is worth the risk of causing the patient anxiety to warn her about this situation. Further, the physician must determine how to encourage interventions to mitigate the potential lethal harm, take care to avoid bias in executing the care plans suggested, and make a determination as to whether or not the algorithm's indication is trustworthy and actionable.[30] To offer another example of an abuse of AI

[28] For more on the significance of prudence cultivated through experience, see, for example, Paul Scherz, "Risk, Prudence, and Moral Formation in the Laboratory," *Journal of Moral Education* 47, no. 3 (2018): 304–315; Courtois, "Among His Own Kin."

[29] See Rebecca Robbins, "An Experiment in End-of-Life Care: Tapping AI's Cold Calculus to Nudge the Most Human of Conversations," *Promise and Peril: How AI Is Transforming Health Care*, STAT, July 1, 2020, 21–31.

[30] For examples of concerns about biases embedded into health care algorithms, see for example, "Bias and Health Inequity," in *Promise and Peril: How AI Is Transforming Health Care*, STAT (July 1, 2020), 86–125.

that disrupts transparency and limits agency, the Targeted Interventions to Greater Enhance Re-entry (TIGER) program has been used against its designed purpose (stated in its very name) in Louisiana: Prisoners flagged at a high enough risk are denied parole. Its intended function was to assist incarcerated individuals with attaining the greatest possible chance for parole by suggesting resources in line with the ultimate goal of reintegration into society.[31]

These examples highlight the temptation to defer to AI according to a presumption of improved efficiency and accuracy, yet sometimes using AI can actually disadvantage cases that deserve prudential judgment or simply a deference to hope. Because AI operates by tracing commonalities and identifying patterns, it may not detect outliers within a data set or account for qualifications or exceptions unanticipated in the initial design. Preserving human oversight in these systems can help to maintain accountability.

There are special concerns for preserving moral agency with regard to the increasing influence of AI technology on society. All technologies embed society members into a wider structure. This not only involves an economic or financial structure, but also structures of preference, implied ultimate goods, and assumptions about what satisfies the human person. One of the often-overlooked ways that AI implicates people into a wider problematic structure is its reliance on energy, materials, space, and resources that may not involve prudent care for the environment.[32] In fact, care for nature is an integral aspect of care for God's providential plan.[33] In order to respect human agency and our decision-making capacities, we should have access to the information that we need for fulfilling our responsibilities to make moral decisions and to avoid illicit cooperation

[31] Richard A. Webster, "Algorithm Deemed This Nearly Blind 70-Year-Old Prisoner a 'Moderate Risk.' Now He's No Longer Eligible for Parole," *ProPublica*, April 10, 2025, propublica.org/article/tiger-algorithm-louisiana-parole-calvin-alexander.

[32] See, for example, Mahmut Kandemir, "Why AI Uses So Much Energy—And What We Can Do about It," *Penn State Institute of Energy and the Environment*, April 8, 2025.

[33] Benedict XVI, *Caritas in Veritate*, §48.

with activities to which we morally object, such as abuse of others' personal data or of the environment.

In seeking to preserve human agency, there is not only a potential benefit in AI analysis; there is also a benefit for the human person who is able to participate in the action alongside AI. Human participation promotes human understanding of the function and application of AI and thus allows for humans to participate more fully in the actions identified as beneficial for the community to which AI is deployed. It is important to protect individuals' participation in the common good. CST distinguishes between false and authentic development, a distinction that helps clarify the proper relationship of technology, human agency, and the common good. False development pursues economic success and capital as the main indicator of a flourishing society, even without attention to the means by which this financial stability is secured or the lifestyles and goods that are emphasized through the resulting structures. It is important that a society cares for the wellbeing of its community members and the broader creation. As Benedict XVI stressed, "The sharing of goods and resources, from which authentic development proceeds, is not guaranteed by merely technical progress and relationships of utility, but by the potential of love that overcomes evil with Good, opening up the path towards reciprocity of consciences and liberties."[34] People are not to be exploited for economic ends. Rather, the economy should be directed toward the flourishing of a society, which necessarily involves the participation of its persons in its function, toward a furthering of their individual vocations.

Indeed, for humans to participate in acts of genuine love, it is requisite that we understand the value of these acts and the needs to which they respond. To seek integral human development, society must rely on the contributions of many different fields, a diverse array of experiences, and the wisdom of many people collaborating. This allows for a stronger and more harmonious community than one that functions according to a one-dimensional promotion of AI. For Benedict XVI, "The correlation

[34] Benedict XVI, *Caritas in Veritate*, §9.

between [integral human development's] multiple elements requires a commitment to *foster the interaction of the different levels of human knowledge* in order to promote the authentic development of peoples."[35] These different levels of human knowledge are needed not only for the sake of thoroughness but also for the sake of permitting a society's diverse array of people to contribute to the common good, leaving their own marks on the work of their community together.

To accomplish this vision of the common good, in which free human agency participates with a depth of love and firmness of insight, it is important to uphold the significance of knowing and loving for human action. This is an insight celebrated in *Caritas in Veritate,* which, as we mentioned earlier, reminds us that "without truth, charity degenerates into sentimentality. Love becomes an empty shell, to be filled in an arbitrary way."[36] Likewise, truth needs to be actualized in human action if it is to have its proper impact on bringing the world into order and peace: "Truth needs to be sought, found and expressed within the 'economy' of charity, but charity in its turn needs to be understood, confirmed and practised in the light of truth."[37] To allow for transparency helps us to act in truth. If we can understand the functionality of AI and hold developers and users of AI accountable for their use of data, impacts on lifestyles, and interconnections with structural ills, this will provide us with the ability to practice charity in the world. This would mean ensuring technology is ultimately used for loving purposes, ones that serve the good of the human community and reflect God's own care for His creation.

[35] Benedict XVI, *Caritas in Veritate*, §30, italics original.
[36] Benedict XVI, *Caritas in Veritate*, §3.
[37] Benedict XVI, *Caritas in Veritate*, §2.

Chapter 11

Positive Visions for AI Design and Distribution

While much of this book considers political questions, in the face of AI, merely political action will not be enough to protect human freedom and agency. A merely regulatory approach can present an antagonistic stance toward the technology and potentially stifle its beneficial possibilities. That is why CST has always offered more than just criticism of new technologies; it also offers a positive social vision that can inspire efforts toward new institutions and forms of life. This chapter explores more constructive approaches to the question of AI's effects on human agency inspired by CST. It examines ways that AI can *support* human agency. The first section argues that economic freedom, and in particular economic subsidiarity, are necessary to develop a better mode of AI. It expands on the discussion of subsidiarity found in chapters 6 and 9 and explores ways to foster human agency through decentralization. The second section presents paradigms of AI design that explicitly aim to support human agency by balancing automation and control in thoughtful ways. These approaches could address some of the problems Part II raised.

Subsidiarity as a Response to AI's Economic Centralization

There is a constant ethical play between centralization and decentralization in political, economic, and other spheres of life. Centralization solves some problems, such as large-scale coordination for addressing society's hardest

challenges, while creating other problems, such as disempowering local leaders and facilitating potentially large-scale human rights abuses. Decentralization also solves some problems while causing others; it might mean the development of local leadership talent and potentially the inhibition of large-scale human rights abuses, while sometimes making it harder to coordinate large numbers of people for the sake of the common good. The ethical component of this balancing act is significant: Successes and mistakes in centralized systems spread widely, while in decentralized systems, effects do not spread as far, but more diversity reigns in both success and failure.

We can think of decentralization in terms of almost any social form. Here we consider the decentralization of political, technological, and economic power. Democracy is a decentralized political power structure. In a democracy, the people work together to find the best path forward, using their own best judgment. There is a saying that if all humans were angels, no government would be necessary, and that if all humans were devils, no government would be possible.[1] But we are neither extreme, so government is possible and necessary, but difficult. This statement is also true for economics, technological development, and so many other things in life. Given the necessity, possibility, and difficulty of human social existence, some principles can provide guidance.

Decentralized technological power might entail something like the right to invent or the right to repair,[2] and, with a more corporate focus, might manifest in something like a startup economy. Decentralized technological power manifests as a feeling of empowerment that we can do

[1] A briefer version of this—"If men were angels, no government would be necessary"—is from "Publius" (pseudonym for James Madison and Alexander Hamilton), *Federalist Papers*, No. 51, February 8, 1788, avalon.law.yale.edu/18th_century/fed51.asp.

[2] See, e.g., Andrew W. Torrance and Eric von Hippel, "The Right to Innovate," *Michigan State Law Review* 793 (2015); "We Need Right To Repair," *Repair.org*, repair.org/; Nicholas A. Mirr, "Defending the Right to Repair: An Argument for Federal Legislation Guaranteeing the Right to Repair," *Iowa Law Review* 105 (2019–20): 2393.

something in terms of creating new things, whether as an individual fixing something at home or as a member of a team working together to innovate.

Decentralized economic power looks something like widespread economic opportunity, where people have agency to freely choose where to work and what to do with their lives. Some of the world has this sort of economic freedom, but not all. Certainly, its extent varies greatly from context to context, especially in relation to economic vitality. If an economy is doing well, there is more freedom for people to choose between careers, products to buy, and other things to do. When the economy is not doing well, economic agency seems to shrink, as people involuntarily lose their jobs, lose the ability to purchase products, and generally become less free. In this case, governments sometimes attempt to stimulate the economy to help restore freedom to their people.

The Principle of Subsidiarity

As discussed in chapter 9, the ethical principle of subsidiarity is the aspect of CST that prioritizes maintaining a decentralized system to the greatest extent possible, while also acknowledging that centralization is sometimes necessary, in the form of aid given from higher to lower levels.[3] In his 1931 encyclical letter *Quadragesimo Anno* Pope Pius XI explained:

> Just as it is gravely wrong to take from individuals what they can accomplish by their own initiative and industry and give it to the community, so also it is an injustice and at the same time a grave evil and disturbance of right order to assign to a greater and higher association what lesser and subordinate organizations can do. For every social activity ought of its very nature to furnish help to the members of the body social, and never destroy and absorb them.[4]

[3] Pontifical Council for Justice and Peace, *Compendium of the Social Doctrine of the Church*, Libreria Editrice Vaticana & USCCB, 2007, §81–82.

[4] Pius XI, *Quadragesimo Anno*, §79, quoted in *Compendium of the Social Doctrine of the Church*, §81.

In other words, people should be enabled and encouraged to solve their own problems, and the more immediately affected any one of us is by something, the more directly that person should be involved in the solution. Groups of people should not take this away from individuals, and larger groups of people should not take it away from smaller groups.

Noting the year he was writing, we should recall the Soviet Communist collectivization schemes from the 1920s and 1930s that were the targets of the Pope's criticism.[5] These schemes not only disempowered individuals and local leaders but were also immensely economically destructive and caused horrendous humanitarian disasters. *Quadragesimo Anno* takes aim at communism in no uncertain terms, excoriating its ideology, implementation, and intentional degradation of human dignity. Communism, perhaps the most centralizing economic system ever, grievously violates subsidiarity. As technology and the economy change, particularly with AI, both old and new forms of centralization should be opposed.

Decentralization can also be a problem if, for example, a government is incapable of providing help and control to its people when needed. However, in our era of expanding technological power, governments are typically not short on means of control and often have many options awaiting use, though limited by legal restrictions. Governments typically do not need to be told to acquire more power, but rather the opposite. So, while both extreme centralization and decentralization can be problematic, here we will mostly discuss the problem of centralization.

Subsidiarity and Universal Basic Income

As we have examined throughout this book, the centralizing power of AI has serious risks associated with it. New technologies always empower differentially: starting with a few, moving to some, then, as time passes, possibly to most people. Empowerment is not necessarily widespread at

[5] Pius XI was also concerned with statist forms of corporatist power emerging in the Italian fascist regime.

the start and sometimes neither at the end.[6] On the other hand, if properly developed, AI could be employed to empower weaker parties to protect themselves against exploitation and control.

One proposal for protecting agency and decentralizing power is the idea of a universal basic income (UBI). Pope Francis spoke positively of UBI several times, saying: "This may be the time to consider a universal basic wage which would acknowledge and dignify the noble, essential tasks you carry out."[7] In this conception, UBI might be more like a wage than a handout, reminiscent of proposals such as the "social investment stipend," which rewards "*socially beneficial* activities the same way we currently reward *economically productive* activities."[8] This could include currently vital but often unremunerated work like childcare and eldercare, volunteering to serve the community, or learning new skills.

---

[6] The anti-subsidiary nature of current AI empowers some (often the already powerful) and disempowers others (often the already less powerful). In its extreme form, AI might empower only a few, while disempowering almost everyone else, or even disempower everyone, including its own creators, who might become locked into a surveillance state or another system of strongly dependent and controlled relationships. Although no one may want such a negative-sum situation, it is not impossible given current trends.

[7] This was in a message to a group of workers, some of whom had lost their livelihoods due to COVID lockdowns: Pope Francis, "Letter of His Holiness Pope Francis to the Popular Movements," April 12, 2020, vatican.va/content/francesco/en/letters/2020/documents/papa-francesco_20200412_lettera-movimentipopolari.html. See also Pope Francis, "Meeting of Popular Movements Promoted by the Dicastery for Integral Human Development," *L'Osservatore Romano, Edizione Quotidiana*, September 24, 2024, vatican.va/content/francesco/it/speeches/2024/september/documents/20240920-movimenti-popolari.html. See also Kate Ward, "Does Catholic Social Teaching Support a Universal Basic Income? According to Catholic Social Thought, Every Human Has the Right to Attain Their Basic Needs," *US Catholic*, April 13, 2020, uscatholic.org/articles/202004/does-catholic-social-teaching-support-universal-basic-income-32025, and Élio Gasda, "Artificial Intelligence: The Future of Labour and Employment," *Asian Horizons* 14, no. 3 (2020): 644–657. dvkjournals.in/index.php/ah/article/view/3206.

[8] Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (Houghton Mifflin Harcourt, 2018), 220–221. See also Kai-Fu Lee, "A Human Blueprint for AI Coexistence," in *Robotics, AI, and Humanity: Science, Ethics, and Policy*, ed. Joachim von Braun, Margaret S. Archer, Gregory M. Reichberg, et al. (Springer, 2021), 261–269.

These proposals not only acknowledge the necessity and dignity of this work, but also help show that there is essentially an *infinite* amount of work in the world. What may be finite is *paid* work, which AI companies have an incentive to vacuum up specifically because it is paid. And we should be clear: lightening the burden of some kinds of work through automation might actually be beneficial because it would allow people more time to do other socially beneficial things that are currently neglected due to time constraints. Caring for others and for the community is sometimes paid, but very often unpaid and done out of love. Much work is done simply for the sake of love, and love is an infinite work.

The specifics of such proposals are quite important. While UBI might seem empowering, it threatens to be disempowering if it makes the recipients dependent upon the UBI source while at the same time giving them no meaningful way to spend their time—essentially reinforcing to the recipient that they are useless.[9] To prevent this disempowerment, the principle of subsidiarity can be an ethical guide: Rather than giving the mere products of production, the means of production themselves should be given, thus providing freedom and motivation to the recipients to support themselves. Additionally, UBI should not be used as an excuse by the powerful to justify the morality of eliminating employment through automation, nor for the extreme concentration of capital, whether in government or private hands.

[9] One possible future is an oligarchy of AI billionaires (or trillionaires), where the disempowered masses are, as in ancient Rome, kept happy with "bread and circuses" (Juvenal, *Satire* 10, v. 81.), e.g., with basic income and AI-powered entertainment. Other futures could be worse, e.g., currently, every nation relies upon human labor to make their country function, but if, due to AI and robotics, human labor becomes less economically valuable, we can imagine the disempowered masses would become mere "liabilities" instead of "assets" because their labor is worthless and all they do is consume resources. Governments might then lose their incentive to protect these people. This attitude, not uncommon in history, reflects a "throwaway culture" (Francis, *Laudato Si'*, §16, 20–22, 43, and *Fratelli Tutti*, §18–21, 188) and "culture of death" (John Paul II, *Evangelium Vitae*, §21, 28, 50, 77, 82, 86–87, 92, 95, 98, 100).

The Beginnings of a Response

The authors of this book are certainly not the first to notice these worrisome trends in concentration of power due to technology. Indeed, many have noticed these concerning trends and have been fighting them—politically, economically, and technologically—for decades. As just one example, computer "hacker culture," while sometimes disparaged as libertarian, anarchic, and anti-authority, also has a protective side that can align with the principle of subsidiarity. Hacker culture can be credited with many innovations and efforts to protect against centralization.

Here we will list a few of the other current strategies to counteract technological and/or economic centralization, and potentially facilitate subsidiarity, thus protecting individual and local freedom.

Distributed ledger / blockchain: One of the most well-known anti-centralization technologies is distributed ledger technologies, such as blockchain, which allow for systems that share information, control, and benefits without a centralized authority. Through various techniques for information sharing and control over how interactions can occur, these technologies permit cooperation in the low-trust and no-trust environments where no central authority exists to enforce control through punishment. While cryptocurrencies are perhaps the most well-known applications for distributed ledger technologies,[10] there are growing opportunities for applications in other fields, such as, for example, information provenance, where new pieces of information can be recorded on blockchain, which could be used to discern real from fake.

Open-source / AI model decentralization: The widespread open-source software movement intersects with AI in the form of open AI algorithms, open data sets, and open models. These open resources can serve to distribute technological and economic power in some ways, permitting access, use, and sometimes (depending on the license) even ownership.

[10] Although there is reason to doubt the true decentralization of cryptocurrencies—see Fabio Vighi, "Why Crypto Won't Free You," *Compact*, July 18, 2025, compactmag.com/article/why-crypto-wont-free-you/.

However, only those with the education, skill, and other necessary materials to use these resources are likely to benefit. For the great majority of people in any society—that is, those who are not technically skilled—open-source work still might offer benefits, but more indirectly.

Data decentralization: Open data sets are becoming more available, but they are still not as developed as the more closed, proprietary data sets that some of the truly massive AI models use. Nevertheless, these smaller and open data sets permit individuals and small groups to at least begin to have the data needed to work on AI as individuals or small groups. Of course, such open data sets can raise privacy concerns that need to be addressed.

Computational decentralization: AI cannot operate without the computational capacity necessary to run the very intensive algorithms for training and inference. Whether computation would be best decentralized through physical means, as in smaller and more distributed data centers, or through means of access to large data centers, is a prudential decision. Computation is extremely costly both in terms of computer chips (and other hardware) and energy, so the question of how to pay for this access is no minor issue. Perhaps the best strategy along this front is the development of more efficient models and smaller, specialized AI applications that require less power.

Talent and skill decentralization: Right now, talented AI engineers are fetching compensation packages previously only seen by professional athletes, with one AI researcher reportedly accepting an offer of $250 million to work for Meta for four years.[11] Top AI companies are buying as much talent as they can afford and centralizing that talent, hoping they will form a team that can outcompete others. Needless to say, most companies cannot afford such extravagant salaries, and those of us without the technical AI know-how could never attract such offers. This centralization of talent and skill reflects the free market and human choice,

[11] Mike Isaac, Eli Tan, and Cade Metz, "A.I. Researchers Are Negotiating $250 Million Pay Packages. Just Like N.B.A. Stars," *The New York Times*, July 31, 2025, nytimes.com/2025/07/31/technology/ai-researchers-nba-stars.html.

and in this way is an expression of human agency, but most people are unable to engage. However, education is now clearly extremely important for helping to develop more widespread talent, which may at some point allow for greater decentralization.

Corporate ownership decentralization: There are many options for distributing the benefits of AI widely in terms of ownership. Traditional ownership of shares of stock is one model that distributes ownership of companies (though this tends to be concentrated, it could be more distributed), and there are many other options, including cooperative ownership, public benefit companies, and nonprofit organization approaches. Indeed, OpenAI has a hybrid structure where a nonprofit organization is a shareholder of a public benefit company,[12] and Anthropic is a public benefit corporation,[13] so these alternatives are already being tried. Perhaps a future phase might see companies actively attempting to distribute shares more widely.

The above are all methods for distributing not just the profits from production (as UBI might), but the means of production, a crucial distinction that preserves local agency. When the means of production are distributed, the necessary dependencies in society become less centralized, less subject to authoritarian fiat, and less fragile. They become a network of relationships, not a central hub with spokes. The agency of the individual is enhanced by broad access to resources, including shared ownership itself. These approaches are not unrealistic. Some, like open-source development, are already well-institutionalized. Other paths for decentralization are not as easy, but that does not mean they are not possible. At this time, we do not advocate for specific policies, but rather state that this overall goal seems both possible and preferable from the perspective of subsidiarity and protecting human freedom.

---

[12] Sam Altman, "Evolving OpenAI's Structure," May 5, 2025, openai.com/index/evolving-our-structure/.

[13] Anthropic, "Making AI Systems You Can Rely On: Governance," 2025, anthropic .com/company.

Possible Catholic Approaches

Building more deeply on the idea of subsidiarity, Catholic ideas on the decentralization of economic power have sometimes coalesced into a system known as "distributism," which seeks to avoid the extreme concentration of the means of production in either private hands (as in unfettered capitalism) or public hands (as in communism). Distributism seeks to *distribute* the means of production widely, thus empowering individuals, small groups, and local communities. If capitalism and communism concentrate the means of production in private and public hands, respectively, distributism would deconcentrate it, placing it in individual hands in the broadest and most inclusive sense possible. Twentieth-century authors such as Hilaire Belloc and G. K. Chesterton were advocates of distributism, and the movement remains alive today, though small.[14] E. F. Schumacher's 1975 book *Small Is Beautiful* also shows similar ideas,[15] and CST has had a constant refrain about developing a world that is more humane and respectful of dignity, which has been inspirational to distributist ideals.[16]

Some might think that this emphasis on individuals and economics sounds like unfettered libertarianism or a capitalist free-for-all, but this would be a misunderstanding. The common good is still the focus, and responsibility toward our fellow people is still a prime task for all, but this focus is achieved by protecting individual rights and the rights of small organizations. Efforts must be made to make sure that people are not economically disempowered, so the role of the government is to protect "the little guy."

---

[14] Hilaire Belloc, *The Servile State* (T. N. Foulis, 1912); G. K. Chesterton, *The Outline of Sanity* (Dodd, Mead & Co., 1927); John Medaille, *Toward a Truly Free Market: A Distributist Perspective on the Role of Government, Taxes, Health Care, Deficits, and More* (ISI Books, 2011); Alexander W. Salter, *The Political Economy of Distributism: Property, Liberty, and the Common Good* (The Catholic University of America Press, 2023).

[15] E. F. Schumacher, *Small Is Beautiful: Economics as If People Mattered* (Harper & Row, 1975).

[16] See, e.g., Pope Leo XIII, *Rerum Novarum*, § 4–6, 8, 11, 13, 46, 47.

Others might accuse such a system of being crypto-communist, full of government interference, attacking large corporations, and "distributing" wealth. This is also a misunderstanding because government is actually reduced by dispersing its power through subsidiarity in government. Corporations can be a part of this system, especially if their ownership is widely held. Finally, the system is not distributing wealth itself but rather the means by which individuals can produce their own wealth through their own effort. Unlike in communism, which removes economic incentives to work, effectively only motivating work through fear, in this alternative view, individual effort is absolutely necessary and a variety of economic incentives remain active.

Economic thinking inspired by CST proposes a third way, which is fundamentally different from Marxist communism and unregulated capitalism. This book is not advocating distributism, merely mentioning it as one possible model for how decentralization of the economy to protect human agency and freedom might be examined. The more distributed that power is, the less liable it is to abuse, and the more empowered individuals and local groups of people are. It is economics at a human scale yet still integrated with the giant world in which we live.

How does this relate to AI? Regarding AI, the above measures to promote decentralization can help distribute security and trust, open products, algorithms, data, computational capacity, skills, ownership, and so on. No doubt there are other ways to make AI more subsidiary. But if we want to achieve a future where there are smaller and more accessible AI models, models aimed at helping the problems of the poor, and models under more local or individual control, inventive people should continue to think, and the rest of us continue to encourage.

What about big AI corporations? The measures against centralization might look like they are putting a bullseye right on the big tech companies, looking to break them up or otherwise handicap them. But this need not be the case. Big tech companies can remain intact if, for example, their ownership structures are adjusted. Perhaps they might become public

benefit corporations, co-ops, or employee-owned companies. They might also just distribute shares more widely or implement other ideas that have yet to be considered.

As an example of a large corporation with a different structure, the Mondragón corporation of Spain began in the 1950s as a company inspired by CST.[17] Today Mondragón exists as a collection of 81 co-ops with 70,000 employees selling products in more than 150 countries, and it still emphasizes its cooperative structure and social responsibility.[18] The Mondragón corporation is, of course, not a big AI company. But the fact that they have, for nearly seventy years, been able to navigate the world with a very different business structure makes it clear that different approaches are possible.[19]

As another example, as we have already suggested, a truly empowering UBI might distribute not just the results of production, but widespread ownership of the means of production themselves. This distribution would not be to a vague "public" that owns things (i.e., the state representing the people, as in socialism and communism), but rather it should be a distribution of the means of production as private property to individuals. This might involve, as a purely hypothetical example, distribution of stock in AI companies to individuals in society, to make them smallholding owners of the AI company. Individuals might then receive a "UBI-like" form of income resembling a stock dividend. To maintain distribution, some restrictions might be necessary; for example, people might not be allowed to sell this kind of stock. Yet while these restrictions could seem strange, this is not socialist or communist national ownership; it is distributed private ownership, with no government involved except in the laws that shape the system.

---

[17] "All Our History," Mondragon, 2025, mondragon-corporation.com/en/history/.

[18] "About Us," Mondragon, 2025, mondragon-corporation.com/en/about-us/.

[19] See, e.g., Xabier Barandiaran and Javier Lezaun, "The Mondragón Experience," in *The Oxford Handbook of Mutual, Co-operative, and Co-Owned Business* (Oxford University Press, 2017), 279–295.

The ideas presented here are not all consistent, and some might even be contradictory. The point is merely to show that greater economic subsidiarity is possible and desirable, and that CST can help this discussion. Whatever the solution to AI centralization might be, it will need some ethical guidance to prevent the worst future possibilities. Subsidiarity is one principle that, if incorporated properly, could benefit society and the individual. The future may be economically quite different. However, for all the challenges we might face, there will always be solutions that can respect human dignity and protect human agency. These solutions will require political will for implementation. But by increasing subsidiarity in the economy, and more broadly sharing the benefits of AI, we could create a path toward widespread prosperity.

Human-Centered Design

A positive vision of AI requires not only shifting the social organization behind the technology or its use. It also requires rethinking design frameworks so that they end up supporting, rather than undermining, human dignity. Protecting human dignity cannot be merely a late addition or protective step added on to a generally pernicious technology. From the first stages of development, the technology should aim to support human agency and virtue. A positive development in this regard is the rise of the framework of human-centered design. This section will first review some of the dangers of dehumanization arising out of current models of AI implementation, as we have discussed in previous chapters, before showing how human-centered design can address these dangers.

Dehumanization and Technology

Human-centered design aims at developing, rather than diminishing, the human person, addressing some of the problems of technical and moral deskilling discussed in chapter 6. Historically, in the modern age, the development from tools to machines has sidelined the human agent: Tools are operated by the user, whereas machines run on their own, without

requiring human input or skill; machines are "autonomous tools."[20] Consequently, human skill and decision has been limited to the beginning of the process, or (currently with LLMs) to its end (editing, selecting results).[21] But the skill to "do it ourselves" is increasingly lost with every new step in automation or every new app.

Not only is our greater dependence on such automated tools dangerous in cases of a system's failure, but it can also diminish us as responsible agents and moral decision-makers. Take, for example, the introduction of self-driving cars: It will predictably and swiftly lead to a loss of driving skills (like the loss of map-reading skills and general orientation due to GPS, or before that, the loss of most people's ability to handle horses). Hopes may be expressed that it will also lead to a reduction in traffic accidents, because, it is claimed, the problem is always the *human* driver. Yet, assuming for the moment that human beings are the problem, we should ask whether eliminating ourselves from the driver's seat is something we should hope for, or whether there might not be something important about the human being's ability to be a problem. The skill of driving a car gives us greater freedom and empowers us as agents. It also makes us dangerous, because we can do serious harm to other participants in traffic. This means that it is a technical skill that requires us to develop a moral skill: Where better could we learn patience, courtesy, controlling rage, being kind in letting someone in, paying attention to pedestrians, and caring for others? It is at the same time a skill shared with others, a communal skill and a responsibility that allows us to practice civic virtue.[22] Automation takes all this away at once. As we examined in chapter 6, it is not only a form of

[20] *Selbständige Werkzeuge*, as discussed in Martin Heidegger, "Die Frage nach der Technik," in Martin Heidegger, *Die Technik und die Kehre* (Neske, 1962), 18. It is a "tool that works by itself," "separated from the idea of the person who practices it"; José Ortega y Gasset, "Meditación de la Tècnica" [1939], translated by Helene Weyl as "Man the Technician," in *History as a System and Other Essays Towards a Philosophy of History* (W. W. Norton, 1941), 148.

[21] Lewis Mumford, *Art and Technics* (1952), repr. Columbia University Press (2000), 82.

[22] Crawford, *Why We Drive*, 241–265.

technical but also *moral* deskilling. Where will we learn responsibility now? Parents know about the importance of giving children increasing responsibilities by giving them more and more power, which is always also the power to do harm. Hence, they give them control and responsibility over pets before entrusting them with younger siblings. Likewise, every new technical skill gives us power, and with this power, responsibility and an opportunity to learn moral skills (i.e., virtues).

Wrestling with reality and its "negative affordances" is a source of skill, but also of our humanity. It forces us to recognize our dependence on and vulnerability to the world and others. As negative affordances are the source for the development of good affordances that give us new access to the world and reality, their atrophy means that our world atrophies as well: Virtual reality may seem expansive, but it is humanly small.[23] Automation that replaces human skill will tend to make our world small by taking away friction with reality.

Further, our decisions depend on choosing which of our wishes and desires to pursue. Such choices and the habits we form around them are part of what makes us mature as human beings. But algorithms are trained to know our desires better than we do, because they are integrated into frameworks of surveillance capitalism to help sell products catering to these desires. If they really know us better than we do, why not automate the choice as well? Why not allow Amazon not only to *suggest* other products, but also to *order* them automatically in the "Internet of Things?" After all, it saves us the mouse click.[24] In this scenario, not only do we cease making decisions with regard to our desires. Instead, the desires are satisfied before we can even become aware of them. But if we are not aware of something, we cannot make a decision; we cannot choose what

[23] Matthew B. Crawford, "Virtual Reality as Moral Ideal," *The New Atlantis* 44 (2015): 28–36.

[24] Something similar is suggested in Crawford, "Virtual Reality as Moral Ideal."

we do not know.[25] Choice is an act of our free will, a faculty which, as we discussed in chapter 2, is one of the things we mean when we say we are made in the image and likeness of God. In this scenario it would be systematically short-circuited and become idle. Not only our technical or moral skill, but our very "muscle" of decision-making, would also be eliminated.

While this eliminates our choices at the *beginning* of automated processes, possible choices at the *end* of such processes are also affected. Processes whereby AI automatically generates possible solutions, but humans select among the results or edit the output, include those increasingly used for the creation of *art*. AI can relieve the artist from the anxiety of sitting in front of an empty canvas or sheet of paper, suffering from writer's block or not knowing what to make in the first place. But perhaps it is important for the artist to experience this anxiety as a source of a new and creative insight. AI, after all, only plagiarizes and remixes already existing patterns, and what it produces is not the expression of anyone's real experience. Is this what the artist is aiming at? And does this process leave even the ability to select and edit unaffected? If we never have to get to the point of a creative intuition ourselves, will we still have the skill to judge the results?

Similarly, in education, learning to express a topic is an important part of learning the topic itself (and hence teaching is the best way of learning). Writing about a topic helps students to learn, so both writing and knowledge will be lost at once. If we let ChatGPT write our texts, will we still be able to tell a good from a bad text once we have lost the skill of writing ourselves? Students will not even be able to learn to judge whether an AI-generated version of an essay is good.

While such uses of technology may be dehumanizing, they conversely tend to humanize the technology itself, perhaps even ascribing personhood to the AI agents that deliver us from skill and decision-

[25] Cf. "one cannot love what one does not know"; Augustine, *De Trinitate* 10.1.3–2.4, and Thomas Aquinas, *ST* I-II, q. 27, a. 2.

making. "Autonomy" used to be a term that, especially since Kant, was employed to understand our own human dignity and freedom. Yet now people speak of "autonomous" cars or drones—that is, it is the machine that is autonomous, not us. For Kant, autonomy was the ability to freely determine the ends of our actions. This, too, may now be relegated to machines that shape our behaviors and desires. They stop being tools; they cease to be means to an end but start to shape the ends themselves.[26] They start to be thought of as agents in their own right, hence the talk of chatbots as "AI agents."

Yet it is still we who are their makers, and we who are made in the image and likeness of God, endowed with free will, determining the ends of technology. Pope Francis was concerned that our age "risks becoming rich in technology and poor in humanity,"[27] and suggested that "artificial intelligence ought to serve our best human potential and our highest aspirations, not compete with them."[28] The threat of this human diminishment is not limited to this or that technical or even moral skill. It is about our status as agents, and as persons who do not outsource their very ability to engage reality consciously and relate to the world responsibly.

Design Principles for More Human-Centered AI

The movement known as human-centered design suggests that more careful programming principles can address many of these problems. For

[26] Ortega y Gasset rightly observes that the dominances of technology, i.e., means, cannot give us ends: "It may well be that one of the basic diseases of our time is a crisis of wishing and that for this reason all our fabulous technical achievements seem to be of no use whatever." "Man the Technician," 121. Yet, contrary to Ortega y Gasset, this is now the role that is at least implicitly assumed by AI as well.

[27] Francis, "Message for the 58th World Day of Social Communications," January 24, 2024, press.vatican.va/content/salastampa/en/bollettino/pubblico/2024/01/24/240124b.html.

[28] Francis, "57th World Day of Peace 2024: Artificial Intelligence and Peace," January 1, 2024, vatican.va/content/francesco/en/messages/peace/documents/20231208-messaggio-57giornatamondiale-pace2024.html.

example, this last concern, of chatbots that mimic humans, suggests a first principle of AI design: Programmers should resist attempts to anthropomorphize the machine. As an expert in human-computer interaction, Ben Shneiderman, argues, "Successful robots utilize the distinctive features of machines. Robots will become more tool-like, tele-operated, and under human supervisory control through well-designed user interfaces that avoid human-like features."[29] AI applications should not try to appear like another human person. Many useful AI applications already avoid such a mistake: search algorithms, diagnostic AIs, autonomous drones, etc. The rise of generative AI, however, has increased the temptation for companies to create AI replicas of humans. The use of AI for companionship and therapy is one of the fastest growing areas of AI development.[30]

Such anthropomorphism raises many ethical and practical challenges. First, it threatens transparency. The first instinct of many companies would be to deploy chatbots as pseudo-human agents to fool customers. Even if the artificial nature of these chatbots is indicated, people are primed to attribute purpose and agency to things with which they converse, leading to practical confusion.[31] Second, such AI increases the danger of manipulation. People might be more likely to want to please or be polite to a pseudo-human bot. This AI can more effectively use emotional levers to shape the user's will.[32] Third, it increases the danger of the replacement of human workers if it seems like a chatbot is equivalent to a human.

---

[29] Ben Shneiderman, "Human-Centered Artificial Intelligence: Three Fresh Ideas," *AIS Transactions on Human-Computer Interaction* 12, no. 3 (2020): 113, doi.org/10.17705/1thci.00131. This section draws extensively on Shneiderman's ideas.

[30] Marc Zao-Sanders, "How People Are Really Using Gen AI in 2025," *Harvard Business Review*, April 9, 2025, hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025.

[31] Sherry Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other* (Basic Books, 2011).

[32] Sam Altman even suggested that processing politeness might cost tens of millions of dollars. See Theo Burman, "Please and Thank You: What Does It Cost to Be Polite to ChatGPT?" *Newsweek*, May 6, 2025, newsweek.com/chatgpt-please-thank-you-cost-2067329.

Even if human workers were not replaced and AI was not anthropomorphic, AI applications can still deskill workers. Therefore, AI applications should be designed to complement rather than replace human skills. They should augment tasks that require virtue-related skills such as those listed in chapter 6: judgment, creativity, empathy, and practical wisdom.

The drive to prevent deskilling does not mean that there is no place for automation. The ethical trick is to balance automation and control in ways that appropriately achieve human ends, allowing high levels of each. Ben Shneiderman uses the example of devices that allow patient control of pain relief. They allow a high level of patient control in the sense that patients guide when to dispense medication. Yet they have automated features that prevent excesses and enable clinician monitoring to prevent the negative effects that can come from the clinical use of opioids.[33]

The danger is that the greater the automation, the more that AI will shape the ends of the user. This is not always bad, because sometimes the designer will want to program ends to protect the user, as with the pain medication dispenser. It is something about which to be concerned, though. It is in setting and shaping ends that people most fully exercise their agency. Moreover, those abilities come through practice and can be lost when not practiced.

Finally, the goal of technology should not be to remove all difficulty, labor, and friction from the world. Technology should support human excellence. Albert Borgmann describes one of the problems of modern technology as its tendency to reduce all goods to commodities.[34] Rather than a dinner being made from local ingredients through skilled preparation, food is delivered through an app; rather than entertainment coming through board games, conversations, or playing music with friends, it comes from a playlist. Our vision is reduced to the technocratic

[33] Shneiderman, "Human-Centered Artificial Intelligence," 116.

[34] Albert Borgmann, *Technology and the Character of Contemporary Life: A Philosophical Inquiry* (University of Chicago Press, 1987).

paradigm, in which everything is given to immediate consumption according to our desires, available at the touch of a button. Yet in so doing, people can lose their skillful, embodied engagement with the world.[35] We are reduced to consumers rather than agents, to having rather than being. It need not be so. AI can introduce us to recipes that match the ingredients we have on hand, making it easier to cook. People can use AI in their creative endeavors. The outcome in part depends on what the AI is designed to do.

Let us take two examples of these principles in action. Alphafold is a tremendously powerful tool for predicting protein structure and designing new proteins. Predicting protein structure is a very difficult task. While there are some principles that aid prediction, these mostly hold only for small stretches of amino acids. Once proteins get big enough, it becomes impossible for people to predict the overall structure from just the amino acid sequence. Thus, determining structures has been largely an empirical process, with much trial and error. Yet finding hidden patterns in large amounts of empirical data, like the database of previously solved protein structures, is something AI is extremely good at. This is what Alphafold does. It performs a task at which humans do not and will never excel, thereby allowing scientists to more easily pursue excellence in the difficult and important tasks of seeking knowledge of biology and cures for disease. Alphafold does not try to pretend to be another scientist, nor does it replace researchers. It augments their work, helping them aim for excellence.

A second example is Ambient AI scribes in medicine. The introduction of the electronic medical record (EMR) created a problem for medical practitioners: They had to input information into the computer during patient appointments. This need drew their attention away from patients toward the screen. Both patients and practitioners hated this result. Some medical practices hired human scribes to fill out the EMR during the appointment, but this was expensive, so many did not. LLMs, however,

[35] See AI Research Group, *Encountering Artificial Intelligence,* 209–217.

are excellent at identifying, transcribing, and summarizing speech. Thus, AI companies developed systems that could automatically fill out the EMR based on the conversation between the practitioner and the patient. The Ambient AI does not pretend to be a person; it just operates in the background. There is no particular human excellence to the bureaucratic task of filling out a chart that Ambient AI replaces. It thus supports rather than undermines the practitioner's agency toward intersubjective engagement with the patient. While these programs are not perfect, and there are concerns about issues such as privacy, their basic design supports truly human action.

In the end, designers must ensure that AI programs remain tools. They are always tools of course, but users can be mistaken about that, seeing them as true "agents." The AI then becomes a tool of the company or institution that designs or implements it to manipulate the end user. It is the human who should be that agent selecting and achieving ends and thereby realizing excellence in a flourishing life. AI can be a great help in this task, but only if it retains its proper role. Many resources exist that can help in thinking about how to ethically design technology[36] and how to ethically strengthen the organizations that create technology, including work in collaboration with the Vatican's Dicastery for Culture and Education.[37]

Conclusion

[36] E.g., Shannon Vallor, Brian Patrick Green, and Irina Raicu, *Ethics in Technology Practice*, Markkula Center for Applied Ethics at Santa Clara University, July 2018, scu.edu/ethics-in-technology-practice/.

[37] José Roger Flahaux, Brian Patrick Green, and Ann Gregg Skeet, *Ethics in the Age of Disruptive Technologies: An Operational Roadmap (The ITEC Handbook)*, Markkula Center for Applied Ethics at Santa Clara University, 2023, scu.edu/institute-for-technology-ethics-and-culture/itec-handbook.

This chapter suggests that there are ways to respect human agency through new forms of social organization and technology design. We again emphasize that the modes of decentralization and AI design that we have discussed are not concrete policy recommendations. Instead, they are offered in the spirit of inspiring a creative and constructive response to our current malaise regarding human agency. Actual responses must come from engaged citizens, programmers, businesspeople, and politicians who strive to meet the needs of the day. There is a way forward in which AI fosters human agency, but we must create it by drawing on a correct understanding of human agency.

Chapter 12

Conclusion

Agency, the root capacity to freely deliberate about and choose the actions we undertake, is a central aspect of human nature, as we are created in the image of God. It is human reason and free will that allow the person to freely turn to God in love and in turn share God's love with others. Human agency thus is not merely about the problem-solving, optimization, or maximization of returns that characterize AI, and to which some people would reduce it. These are indeed elements of human agency insofar as people exercise a responsible stewardship for the world and care for others. But most fundamentally, human agency is an openness to relationship, to God through Christ and to others. Our agency enables us to act through charity in truth.[1] In so doing, we realize the end for which we were made. Protecting human agency is thus central to protecting human dignity and human flourishing.

Just as it is ordered toward relationships, human agency is also shaped by our relationships to others and to the world. Most fundamentally, it is shaped by God, who enables all our action through the act of creation and leads us to eternal fulfillment through grace. But it is also shaped by the social and natural environments in which we live: our families, schools, media, workplaces, cultures, and nations. These social surroundings form our abilities to act, by molding our character, giving us the tools and skills

[1] Benedict XVI, *Caritas in Veritate*.

to be effective, and granting us opportunities to act. These influences are often positive and essential for effective human agency.

Yet social situations also can degrade agency by malforming our dispositions, denying the disadvantaged the basic goods that they need for effective agency, or constraining people's ability to act. As we described in the second part of the book, AI has all too often been a tool for degrading human agency. It manipulates us through nudges and other strategies of surveillance capitalism. It eliminates opportunities for people to act well through deskilling or delegating political decisions to algorithms. Degraded information ecosystems cause us to lose contact with the truth upon which we can act. Many of these attacks on agency at the level of individual manipulation, or the destruction of the social ecologies necessary for effective agency, need to be addressed through the kind of regulations we discuss in chapter 10.

Consequences that are perhaps just as damaging occur when we misunderstand what AI is because we misunderstand human agency. AI is a powerful tool that may transform society for the better. But AI applications cannot be an agent in the way humans are. They are ultimately statistical and mathematical artifacts. Lacking intellect, freedom, a recognition of the good, and any grasp of semantic meaning, they cannot act with deliberation. They execute programs and optimize results. In mistaking our AI applications for agents, we risk falling into a form of idolatry. We forget that these programs are the works of our hands and imagine them as more powerful than they actually are.

Hans Urs von Balthasar notes how humans "invent gods for themselves, cosmic or supra-cosmic beings that exhibit themselves majestic and free at a level where men stand in the grip of fate. . . . They put words in the mouths of their gods which correspond to their own dreams and longings; they ascribe to them deeds and powers that surpass their own, [that] manifest their idea of the absolute."[2] When we hand over our agency

[2] Hans Urs von Balthasar, *The Glory of the Lord*, vol. 6, *Theology: The Old Covenant*, trans. Brian McNeil (T&T Clark, 1991), 31–32.

to our own creation, have we not, as von Balthasar accuses, made an idol, with "qualities we would like to have" and to which we can ascribe "deeds and powers" that surpass our own? Yet gods of our own creation always fail, for they are made in our image, rather than God's.

As society has become more shaped by digital technology, it has become easier to misunderstand human agency as a kind of information-processing and to incorrectly see AI as an agent. Since people spend increasing amounts of time online, they are tempted to forget that to be an agent is, ultimately, to act in the world. This requires a body as well as a mind. When we think of AI as an agent that is exactly like humans, we can veer into a form of Gnosticism that despises the body. The transhumanism discussed in chapter 8 is the modern form of this Gnosticism, envisioning people as merely information patterns that can ultimately be uploaded to machines. Such a view has repercussions not only for us, but also for the environment. In the encyclical *Laudato Si'*, Pope Francis inveighed against technologically enhanced sins against nature, which he noted are ultimately also sins against the poor, since "human life is grounded in three fundamental and closely intertwined relationships: with God, with our neighbor and with the earth itself."[3] AI is not well suited for the larger physical environment, as it requires vast quantities of energy and water, most of which, currently, derive from nonrenewable sources. This Gnostic scorn for the body and the world was strongly denounced by the patristic fathers and continues to be rejected by the Church today.[4] The early Christians always stressed the unity of body, mind, and spirit. According to St. Paul, a person is a unified whole. Bultmann writes, "The most comprehensive term which Paul uses to characterize man's existence is *soma*, body."[5] For Paul, "Man does not have a body, he is a body. He is

[3] Francis, *Laudato Si'*, §65.

[4] For the most recent magisterial discussion of Gnosticism, see Congregation for the Doctrine of the Faith, "*Placuit Deo*," 2018, press.vatican.va/content/salastampa/en/bollettino/pubblico/2018/03/01/180301a.html.

[5] Rudolf Bultmann, *Theology of the New Testament* (Baylor University Press, 2007), 192.

flesh-animated-by soul, the whole conceived as a psycho-physical unity."[6] A god created in the hope of transcending our own materiality is the converse of the God of Christianity—a transcendent God who created us and, through the incarnation of Jesus, shared and sanctified our embodied condition and our common earthly home. When we thoughtlessly follow our current path, buying the hype that "someday" AI will act as our agent to solve all our problems, including climate change, while primarily using it as a toy, to solve trivial problems, or to manipulate others, we are forgetting the image of God in which we were created, an image that calls us to be, ourselves, God's hands, in stewardship of and relationship to all of creation.

These reflections on the human person have implications not only for how we understand the nature of AI and our relationship to it, but also for how we use it. In general, our use of AI applications will better support human agency to the degree that they foster human wholeness, connection to our material environment, and genuine human relationships. As chapter 11 described, decentralized AI that supports economic and political subsidiarity also supports human agency. For example, an AI application constructed and used locally for classroom management will likely better support human agency than a university-wide AI application, as it can be more responsive to the immediate concerns of the teachers as they act. An AI built to support the operations of a small business will usually be more conducive to fostering human agency than one built and implemented by a large and centralized corporation. AI applications will better support human agency if they are responsive to the concerns of local communities, precisely because the human person is bodily enmeshed in (and called to be a steward of) a particular environment and particular relationships with others. Of course, AI applications will not become supportive of human agency merely by virtue of being decentralized, since local authorities could also design manipulative applications. Likewise, it is possible that a more large-scale centralized AI application might support

[6] John A. T. Robinson, *The Body: A Study in Pauline Theology* (SCM Press, 1952), 18.

human agency better than a smaller-scale AI. For example, an AI tasked with water distribution at a state or national level may well serve human agency better than a collection of local AIs assigned a similar task. In general, though, AI applications that support subsidiarity will be more likely to promote human agency than larger-scale applications by providing tools and information tailored to a specific situation.

AI applications also better support human agency to the degree that they foster partnerships between AI and humans rather than replacing or undermining human actors and activities, as chapter 11 discussed regarding human-centered design. It is helpful to think in terms of a continuum, with AI applications replacing humans entirely on one end and AI used as a tool on the other. An AI that assists a teacher in an instructional task supports human agency to a greater degree than one that replaces a teacher; an AI that assists a health care worker in arriving at a diagnosis supports human agency to a greater degree than one that makes a diagnosis on its own. Of course, it is possible that an AI might replace some human activities while still supporting agency. Pope Francis spoke favorably, for example, about AI's capacity to provide "liberation from drudgery" by taking over routine or unpleasant activities.[7] In general, however, AI is a human creation, developed to aid rather than to supplant.

As these last paragraphs and chapters suggest, AI is a tool that can be designed and used to foster human agency. Problems occur because of misunderstandings of human agency, machines, and our relation to the world and others. Some of these mistakes become built into the very design of AI applications. Those problematic designs are not a necessary outcome of the development or use of AI. It is possible to create applications that could make great contributions to human flourishing. Yet insofar as we use it carelessly, deploy it manipulatively, or regard it as a surrogate for other humans, we forget the very image of God, in which we are created, an image made to love and serve our Creator, each other, and to care for the beautiful and irreplaceable creation God has given us. In Genesis 1, the

[7] Francis, "Message for the 57th World Day of Peace."

first command God gave to human beings is to exercise stewardship over creation. Our most important, most creative, work is the journey of our personal vocations. The greatest threats to agency are those that prevent us from answering and enacting the particular missions to which we are called by God. Acting in the world with agency is the responsibility we each carry with us along our vocational paths—one we must not, cannot, offload to machines.

CONTRIBUTORS

Nathan Colaner is teaching professor in the Department of Management at Seattle University, where he is also director of business analytics. His research and teaching focus on the ethical implications of the development and implementation of artificial intelligence, cybersecurity, and analytics. He consults for the National Science Foundation, the National Institute of Standards and Technology, the Vatican, and various other business and nonprofit organizations.

Jeremiah Coogan is assistant professor of New Testament and early Christianity at the Jesuit School of Theology in Berkeley, California. For the 2025–26 academic year, he is a member of the Institute for Advanced Study in Princeton.

Mariele Courtois is a theological biomedical ethicist whose research interests include genetic engineering, health care, and technology ethics. She draws from Edith Stein's work on theological anthropology for better understanding the relationship between disability and identity. She is an assistant professor of moral theology and director of the Center for Technology and Human Dignity at Benedictine College in Atchison, Kansas. She earned a PhD with distinction in moral theology and ethics and a master of philosophy in theology and religious studies from the Catholic University of America. She also holds an MTS from the University of Notre Dame and a BS in biology from Loyola Marymount University.

Heather Foucault-Camm is the program director for the Science and Religion Initiative at the McGrath Institute for Church Life at the University of Notre Dame, where she develops pedagogical programming

and lectures on key questions on the relation between science and the Catholic faith. A former high school science teacher and curriculum designer, Heather has a BSc and MSc in chemical physics, a PGCE, an MA in theology. She is pursuing a PhD in moral theology at Notre Dame. Heather's scholarly work focuses on the social implications of emerging technologies, with a particular interest in artificial intelligence. She is a scholar associate of the Society of Catholic Scientists upon invitation from its board and is a Lay Dominican of the Province of St. Albert the Great.

Matthew J. Gaudet is director of ethics programs and initiatives for the School of Engineering at Santa Clara University, where he is tasked with developing and managing a program of moral, philosophical, and theological formation for engineering students, preparing them to be conscious, reflective, and ethical as the future builders and maintainers of a technological society. Gaudet's research lies at the intersection of Catholic and philosophical ethics with the social sciences, especially the application of ethics to the topics of war and peace, the university, disability, and technology as well as the field of ethics education. He is the co-author of *Eight Theories of Justice: Perspectives from Philosophical and Theological Ethics* (Fortress, 2025) and co-editor of three special issues of the *Journal of Moral Theology*, including a 2022 issue on the topic of artificial intelligence.

Brian Patrick Green is the director of technology ethics at the Markkula Center for Applied Ethics at Santa Clara University. His work focuses on AI and ethics, technology ethics in corporations, the ethics of space exploration and use, the ethics of technological manipulation of humans, the ethics of mitigation of and adaptation towards risky emerging technologies, and various aspects of the impact of technology and engineering on human life and society, including the relationship of technology and religion (particularly the Catholic Church). Green is author of the book *Space Ethics* (Rowman & Littlefield, 2021), co-author

of the book *Ethics in the Age of Disruptive Technologies: An Operational Roadmap* (The ITEC Handbook, 2023), and co-author of the Ethics in Technology Practice resources. He is co-editor of the book *Religious Transhumanism and Its Critics* (Bloomsbury, 2024), co-editor of a special issue of the *Journal of Moral Theology* on artificial intelligence, and a contributing author to *Encountering Artificial Intelligence: Ethical and Anthropological Investigations* (Pickwick, 2024).

Noreen Herzfeld is director of Benedictine spirituality and the environment at St. John's School of Theology and Seminary and senior research associate at the Institute for Philosophical and Religious Studies in Koper, Slovenia. She holds degrees in computer science and mathematics from The Pennsylvania State University and a PhD in theology from The Graduate Theological Union, Berkeley. Recent works she has written or co-edited include *The Artifice of Intelligence: Divine and Human Relationship in a Robotic Age* (Fortress, 2023); *In Our Image: Artificial Intelligence and the Human Spirit* (Fortress, 2002); *Technology and Religion: Remaining Human in a Co-Created World* (Templeton, 2009); *Encountering AI: Ethical and Anthropological Investigations* (Pickwick, 2024); *Religious and Cultural Implications of Technology-Mediated Relationships in a Post-Pandemic World* (Bloomsbury, 2023); and a special issue of the journal *Religions*, entitled *Religion and the New Technologies* (2017).

Cory Andrew Labrecque is professor of bioethics and theological ethics and the inaugural chair of educational leadership in the ethics of life at the Faculty of Theology and Religious Studies at Université Laval in Quebec City, where he serves as vice-dean. Cory earned a BSc in anatomy and cell biology, an MA in religious studies specializing in bioethics, and a PhD in religious ethics at McGill. Cory's teaching and research explore how the Abrahamic religions—focusing on the Roman Catholic tradition—approach ethical issues in medicine (especially at the end of life),

biotechnology (with an interest in AI), and the environment. Cory is a corresponding member of the Pontifical Academy for Life, president of the Canadian Bioethics Society, vice-president of the National Committee for Ethics and Ageing (in Quebec), and a member (appointed by the Canadian Conference of Catholic Bishops) of the Faith and Life Sciences Reference Group of the Canadian Council of Churches.

Anselm Ramelow, OP, is Professor of Philosophy at the Dominican School of Philosophy and Theology in Berkeley and a member of the Core Doctoral Faculty at the Graduate Theological Union. Fr. Ramelow also taught at the University of San Francisco and the Munich School of Philosophy and is a senior fellow at the Berkeley Institute. He obtained his doctorate under Robert Spaemann in Munich on Leibniz and the Spanish Jesuits (*Gott, Freiheit, Weltenwahl*, Brill, 1997) and did theological work on George Lindbeck and the question of a Thomist philosophy and theology of language (*Beyond Modernism—George Lindbeck and the Linguistic Turn in Theology*, Ars Una, 2005). He contributed articles to the *Historisches Wörterbuch der Philosophie* and essays on topics at the intersection of philosophy and theology, as well as a translation and commentary on part of Aquinas's *De veritate*. He continues to work on questions of free will, philosophy of religion (miracles, existence, and nature of God) and philosophical aesthetics.

Paul Scherz is the Our Lady of Guadalupe Professor of theology at the University of Notre Dame and a program chair for the Notre Dame-IBM Tech Ethics Lab. Building on a dual training in genetics and moral theology, his work examines the intersection of theology, science, medicine, and technology. He is the author of *The Ethics of Precision Medicine* (University of Notre Dame Press, 2024), *Tomorrow's Troubles: Risk, Anxiety, and Prudence in an Age of Algorithmic Governance* (Georgetown University Press, 2022), and *Science and Christian Ethics*

(Cambridge University Press, 2019). He has previously taught at the Catholic University of America and the University of Virginia.

Margarita Vega is professor of philosophy at the Dominican School of Philosophy and Theology in Berkeley, California. She has also taught at the University of Valladolid (Spain) and the University of California, Berkeley. Her work lies at the intersection of metaphysics, philosophical anthropology, and the philosophy of mind, with a special interest in the ontological implications of artificial intelligence. Combining analytic philosophy with historical and Thomistic methodologies, her research explores questions of personhood, the mind-body relation, the nature of consciousness, and aesthetics. She is the author of *Aristóteles y la metáfora* (Ediciones Universidad de Valladolid, 2004) and of multiple articles on metaphysics and the philosophy of mind, including "Aquinas Walks into the Chinese Room," "A Concausal Approach to the Mind–Body Problem," "Person and Collective Intentionality," "Person and Rationality," and "Once Again, What Counts as Art?" Her current work examines how classical metaphysical concepts can illuminate debates about the self, embodiment, and personhood.

Andrea Vicini, SJ, is chairperson, Michael P. Walsh Professor of bioethics, and Professor of theological ethics in the theology Department at Boston College. Medical doctor and pediatrician (University of Bologna), he is an alumnus of Boston College (STL and PhD) and holds a Sacred Theology Doctorate from the Pontifical Faculty of Theology of Southern Italy (Napoli). He is co-chair of the international network Catholic Theological Ethics in the World Church. His research includes fundamental moral theology, theological bioethics, global health, biotechnologies, and environmental issues. In 2015–16, he was research fellow at the Center of Theological Inquiry (Princeton, NJ) on the societal implications of astrobiology. He authored *Genetica umana e bene comune* (2008), co-authored *Encountering Artificial Intelligence: Ethical and*

Anthropological Investigations (2023), and co-edited: *Plastic Pollution, Theological Ethics, and the Call of* Laudato Si' (2025); *The Rising Global Cancer Pandemic: Health, Ethics, and Social Justice* (2022); *Ethics of Global Public Health: Climate Change, Pollution, and the Health of the Poor* (2021); *Reimagining the Moral Life: On Lisa Sowle Cahill's Contributions to Christian Ethics* (2020); *Building Bridges in Sarajevo: The Plenary Papers of CTEWC 2018* (2019); *Just Sustainability: Technology, Ecology, and Resource Extraction* (2015); and *The Legacy of Vatican II* (2015).

Joseph Vukov is an associate professor of philosophy and the associate director of the Hank Center for the Catholic Intellectual Heritage at Loyola University Chicago. He is the author of several books, including *Staying Human in an Era of Artificial Intelligence* (New City Press, 2024). His writing has also appeared in venues including *The Chicago Tribune*, *America Magazine*, *Public Discourse*, *Fox Opinion*, and many academic journals. Vukov serves as President of Philosophers in Jesuit Education and is the winner of the 2025 St. Ignatius Loyola Award for Excellence in Teaching.