"*Staying Human* is not merely another book about AI — it is a philosophical page-turner, a remarkable blend of brilliance and plain-spoken writing that is just enjoyable to read. The book is both intelligent and fun, a rare combination, and a blessing to any reader thinking about what it means to be human in the era of AI."

Brian Patrick Green, Ph.D.
Director of Technology Ethics,
Markkula Center for Applied Ethics
Adjunct Lecturer, School of Engineering,
Santa Clara University

"In an age and time when conversation concerning 'Artificial Intelligence' has the tendency to devolve into slogans and superficiality, Joseph Vukov's *Staying Human in an Era of Artificial Intelligence* is understandable, readable, and even enjoyable!"

+Bishop Kevin W. Vann
Bishop of the Diocese of Orange, California

"It is not often that an author manages to be simultaneously clear, enjoyable, and wise. Yet in *Staying Human*, Vukov combines these qualities and puts them at our service in understanding, and responding to, the challenge of living a genuinely human life surrounded by AI."

Christopher T. Baglow, Ph.D.
Academic Director, Science & Religion Initiative
McGrath Institute for Church Life,
University of Notre Dame

STAYING HUMAN

In an Era of Artificial Intelligence

MAGENTA
Bold Christian voices healing divides

Living the Feminist Dream
A Faithful Vision for Women in the Church and the World
by Kate Bryan

Keep at it, Riley!
Accompanying my Father through Death into Life
by Noreen Madden McInnes

Rehumanize
A Vision to Secure Human Rights for All
by Aimee Murphy

The Church's Mission in a Polarized World
by Aaron Wessman

The Perils of Perfection
On the Limits and Possibilities of Human Enhancement
by Joseph Vukov

STAYING HUMAN

In an Era of Artificial Intelligence

Joseph Vukov

NCP
NEW CITY PRESS

Published by New City Press
202 Comforter Blvd.,
Hyde Park, NY 12538
www.newcitypress.com

Staying Human in an Era of Artificial Intelligence

Cover and Layout by Miguel Tejerina

ISBN: 978-1-56548-599-0 (Paperback)
ISBN: 978-1-56548-605-8 (E-book)

Library of Congress Control Number: 2024935466

For my parents,
who taught me how to stay human

Contents

Series Preface

Does the book that you are about to read seem unusual? Perhaps even counterintuitive?

Good. The Magenta series wouldn't be doing its job if you felt otherwise.

On the color wheel, magenta lies directly between red and blue. Just so, books in this series do not lie at one limit or another of our hopelessly simplistic, two-dimensional, antagonistic, binary imagination. Often, in the broader culture any answer to a moral or political question gets labeled as liberal or conservative, red or blue. But the Magenta series refuses to play by these shortsighted rules. Magenta will address the complexity of the issues of our day by resisting a framework that unnecessarily pits one idea against another. Magenta refuses to be defined by anything other than a positive vision of the good.

If you understand anything about the Focolare's dialogical-and-faithful mission, it should not surprise you that this series has found a home with the Focolare's New City Press. The ideas in these books, we believe, will spark dialogues that will heal divides and build unity at the very sites of greatest fragmentation and division.

The ideas in Magenta are crucial not only for our fragmented culture, but also for the Church. Our secular idolatry— our simplistic left/right, red/blue imagination—has oozed into the Church as well, disfiguring the Body of Christ with ugly disunity. Such idolatry, it must be said, has muffled the Gospel and crippled the Church, keeping it from being salt and light in a wounded world desperate for unity.

Magenta is not naïve. We realize full well that appealing to dialogue or common ground can be dismissed as a weak-sauce, milquetoast attempt to cloud our vision of the good or reduce it to a mere least common denominator. We know that much dialogic spade work is yet to be done, but that does not keep the vision of the Magenta Series (like the color it bears) from being *bold*. There is nothing half-hearted about it. All our authors have a brilliant, attractive vision of the good.

One of those authors, Joe Vukov, has already established himself in the Magenta Series in precisely this way with his important book *Perils of Perfection*. In this current book, his second for the series, he builds on the approach he first took in accessing transhumanism and applies a similarly compelling approach to artificial intelligence. The Magenta Series wants authors like Vukov who can address complex issues and high level ideas and express them in a Christian key, in an accessible style (often with storytelling at the center), and in ways which resist the "two sides" narrative. Once again, Vukov has marshalled his intellectual chops, engaging style, and faithful Christian commitments in arguing for what is good about AI and against what is problematic. Refreshingly, this approach resists the "AI is the bestest thing ever" vs. "AI is the beginning of the end of the world" binary imagination that tends to shape the current discourse. I invite you to join Vukov once more as he helps us understand the complexity of the issues in such a way that we are unafraid to access the opportunity of AI—while at the same time firmly “staying human” in the face of powerful forces pushing us in very different and very dangerous directions.

Enjoy!

Charles C. Camosy
Series Editor

Introduction

I'm a nerd. And before you ask, no, not what passes for nerdiness today. An authentic, old-school nerd. I watch YouTube reviews of board games and keep a spreadsheet documenting which I'll play next. If I can find other spare time, I read fantasy novels for fun (Or is that cool now?). In high school, there was slim chance you'd find me practicing layups or prepping for prom—but a decent chance you'd find me programming my trusty TI-84 calculator. And I spent a good chunk of the 1990s glued to episodes of *Star Trek: The Next Generation*, taped in grainy resolution on my family's VCR.

In case you haven't seen it, the show features Captain Jean-Luc Picard and the crew of the Starship *Enterprise* as they explore "space, the final frontier." The mission of the *Enterprise*, as Captain Picard puts it in the opening credits, is "to explore strange new worlds; to seek out new life and new civilizations; to boldly go where no one has gone before." And that's exactly what they do, one glorious episode at a time.

Along the way, we encounter snapshots of a future that formed and captivated yours truly. The Holodeck: a virtual reality platform in which crew members could navigate a range of programmed worlds and experiences—a 1960s gangster movie; a James Bond-style tale of intrigue and espionage; Sherlock Holmes's nineteenth-century England. The android Data: the *Enterprise*'s second officer, programmed to replicate human behavior, speech, and communication. The Borg: a group of former individuals incorporated into a hive mind of cyborgs, now roaming the cosmos in search of still more groups to assimilate. Transporters: technology that can disassemble people and objects only to reassemble them in nearby locations.

Some of *Star Trek*'s technologies have yet to arrive on the scene. Transporters, for one, won't be replacing your morning commute any time soon—though that hasn't stopped philosophers from spilling bottles of ink thinking about the implications (Would the reassembled version of me really be *me*? What if the teleporter misfires and creates *two* versions of me? Which one is me and which a shabby copy?). The Borg, too, have stayed fictional—though sometimes our political and social media landscapes share striking similarities.

We're closer to building the Holodeck. No, we're not there yet. But walk into any mall, and you can purchase a pair of VR goggles or an Apple Vision Pro and then log into your choice of experiences. I know because I have done it. I have used a pair of goggles to play "basketball" with a friend across town, competed in paintball with a group of strangers, and attended a virtual reality church with congregants from around the world. The VR and Apple Vision Pro experience is decidedly less immersive than the Holodeck but comes closer to anything we had just a few years ago. We needn't buy a ticket on the *Enterprise* to "seek out new life and new civilizations." New worlds have arrived on Earth, inside a pair of goggles.

We have made even greater strides toward creating someone (or, more accurately, some*thing*) like Data. I'm talking, of course, about artificial intelligence, or AI: the main subject of this book. AI has been with us for years: every time you log in to social media or ask Siri for directions or take Amazon's recommendation for your next book, you're interfacing with an AI. Yet in 2023, with the rise of *generative* AI—that is, AI capable of *generating* novel content—the technology has exploded into our collective imagination. And not only our imaginations, but the dollars of investors, the headlines of newspapers, the hopes of futurists, and the nightmares of plagiarism checkers. As I write these words, we stand on a precipice. College essays; news articles; social media posts; YouTube videos: already, these areas of formerly human influence are being overrun by AI. There's no question anymore: AI will infiltrate our lives in myriad ways. Yet, as of now, we are unsure of what form that infiltration will take. The age of the internet seems passé, old hat, and dull. The era of AI has arrived.

Our collective reaction to the rise of AI has been a mixture of excitement and apprehension. Excitement at the possibilities that lie before us. But apprehension at how those possibilities might change or undermine practices and values we hold dear. In the pages that follow, I will suggest that we should aim to strike a healthy balance between these reactions. AI does indeed contain the potential for misuse. It can amplify existing social problems, chip away at our humanity, and erode our spheres of action and influence. Personal and legal bumpers must be put in place, lest we incorporate AI into our lives in destructive ways. Yet AI needn't be shunned. It provides real possibilities for doing good. Our fear of AI is well-merited. But our excitement about it is well-merited too. The rise of AI, in fact, provides an opportunity—an opportunity for us to reflect on the significance of *what we do and how we do it*. As AI becomes incorporated into our lives, we must reflect on the everyday tasks and hassles that can be ceded to the algorithm, and those that must be retained for ourselves. This kind of reflection will be crucial for guiding us as we adopt AI on a large scale, but perhaps more crucially, for guiding us as we invite AI into the little things of our lives: from personal correspondence to nine-to-five workflows, from fitness tracking to household management.

The rise of AI also provides the opportunity for a different kind of reflection. In particular, it provides the opportunity for us to reflect on *who we are*. Here's a teaser trailer of where we are heading in our discussion. When we view an AI-generated meme, when we dabble in creating AI-generated art, when we craft an essay or email or poem using AI, there seems to be something *missing*. This is the case even (and especially) when the AI succeeds at replicating the real thing. However—despite this feeling of there being something "missing" from the products of AI—according to many ways of understanding human nature, it isn't clear why this would be the case. Isn't mimicking human life good enough? What exactly does AI miss? Why does its "intelligence" strike nearly all of us as merely artificial?

Many views of human nature—or, to use a ten-dollar word, many *anthropologies*—are not up to the task of answering these questions. A focus in this book will therefore be introducing an anthropology that can answer them. I'm a Christian, and the view I'll argue for is grounded in a Christian (and more specifically, Catholic) way of understanding ourselves. But it is also a view that is readily available for others to adopt. According to this view, human beings are embodied yet also ensouled. An algorithm is neither. The upshot? Any attempt made by an AI at replicating human intelligence will inevitably remain artificial. Not the real deal. The view of human nature I will be presenting thus diagnoses those aspects of human nature that new Silicon Valley startups can never replicate. It vindicates our near-universal intuition that something crucial is missing from AI. Along the way, the view puts its finger on what makes us human to begin with.

AI poses a real and present danger. As we will discuss in the pages that follow, it contains the capacity to amplify social problems, drive a wedge further into our already-polarized society, and sow seeds of distrust in communities and personal relationships. It has the capacity to erode spheres of influence and activity that should be retained for ourselves. When approached without a robust sense of who we are, AI also threatens to undermine our self-understanding. To a degree beyond any previous technology, AI can make us forget ourselves. In this new era of AI, we must consciously make a choice: *to stay human*. This book provides a map and the tools for doing just that.

A note on how I have written it: rather than making you slog through sections of uninterrupted type, I have written this book in short, digestible pieces. Chapters you can read before bed, on a lunch break, or during a toddler's nap. Each chapter offers something new and tackles the themes of our discussion from a different angle. I have excised the boring material and left only the juicy bits. Some of the chapters can be read on their own, though the book ultimately does build toward a cohesive argument. So I do recommend reading the chapters in order.

How, then, to stay human in an era of AI? I don't have all the answers. Or even all the right questions. As a philosopher, I believe careful reflection on important topics is valuable even if some questions remain. So if you finish the book with some lingering questions, that's okay. Welcome, even. By the end of the book, though, I hope to have guided you on your journey toward staying human in an era of AI.

Chapter 1

Learning Machines

In late 2022, ChatGPT showed up everywhere: in news headlines, YouTube videos, theological conversations, and the policies of hand-wringing academic administrators. In case you missed it, ChatGPT is a generative artificial intelligence (AI), a so-called large language model (LLM) that generates text that appears to have been written by a human. Users can ask it to compose essays, write code, plan a child's birthday party, or write a haiku. Or, if you have some time, you can simply have a conversation with it. The responses ChatGPT generates are not copy-pasted from somewhere on the Internet, nor are they hastily penned by an English major in Cleveland hustling for rent money. The responses are genuinely novel pieces of text, generated on the fly using some of the most complex algorithms ever created.

The essays it churns out? Solidly B-level. Its poetry? Frankly, not great. In other areas, though, ChatGPT sings. It can churn out boilerplate emails and announcements more efficiently than an entire cubicle suite of mid-level executives. It can produce social-media-ready marketing materials for your startup's new product. It can write a letter of recommendation in seconds rather than hours (If any of *my* students are reading this, don't fret. . . . I wrote *your* letter the old-fashioned way). It can draft a meeting agenda quicker than any administrative assistant. And try asking it to create acronyms, title your new podcast, or write a heartfelt sympathy note. In all these areas and others, ChatGPT crafts prose that (rather embarrassingly) rivals the best efforts of humans working with pen and paper.

The rise of AI doesn't stop with LLMs. OpenAI, the company behind ChatGPT, has developed a host of other products. DALL-E generates images from descriptive text: faux photos; computerized Cubism; plausible post-Impressionism. CLIP goes the other direction, generating text descriptions of images—a caption-writer's killer. Sora produces stunningly realistic video clips. Whisper transcribes speech to text and other languages to English. Jukebox generates novel music. Not just sheet music with notes, but actual tunes. And that's just OpenAI. Rival companies have been launching their own products quicker than ChatGPT can churn out essays for Philosophy 101.

You don't need me to tell you that these new developments will upend life as we know it. In the era of AI, beneath every thank you note in the mailbox will lurk a question: did Cousin Magnus really write this, or did he use an AI? Our social media feeds will become dominated by AI-generated advertisements and correspondence. AI-generated images will become indistinguishable from snapshots of real life. Bot-created video and music and poetry and marketing materials will elbow aside versions labored over by actual humans. Jobs previously thought of as skilled labor will become obsolete, joining the ranks of positions at typewriter and eight-track repair shops.

In each of these areas, a small part of our individual and shared humanity is frittered away. Or eliminated outright. The revolution is upon us. It didn't come in the form of a riot. Or even a Terminator robot. It arrived in the form of an algorithm. If we are to resist this revolution—and to stay human in an era of AI—we must be prepared to resist the incoming tide. The first step: understand how AI works to begin with.

Artificial Intelligence 101

What is artificial intelligence? We could spend a lot of time debating the answer. There isn't an agreed-upon definition. Still, the one provided by the Organisation for Economic Co-Operation and Development puts us in the ballpark of something most scholars would agree on: "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."[1]

Let's unpack that. First, an AI is a machine. Remember that. A machine. Not a human. Second, an AI operates according to objectives, either explicitly or implicitly defined. Put differently: an AI is always built *for a purpose*. Maybe that purpose is to recommend videos you'll like on YouTube. Or maybe it is to craft prose that looks human-generated. Later, we will discuss this aspect of AI in terms of "alignment." Third, an AI makes inferences from inputs to generate outputs. To fulfill the purpose it has been assigned, a fully functional AI takes what we feed it—a prompt or an image or a spreadsheet—to generate a certain kind of output. And finally, the outputs of an AI, which can be used to "influence physical or virtual environments," can take the form of "predictions, content, recommendations, or decisions." Not an exhaustive list, but a good start.

All this, however, is very abstract. Let's bring it back down to earth by working through how a simple AI is built. Start with a task you might need to complete. A relatively easy one. Let's say you have a photo-taking obsession. Your pictures fall neatly into two categories: pictures of your puppy, Jack, and pictures of the mountains surrounding your home. The problem: your phone's memory is full to bursting. And your photo-taking obsession isn't slowing down any time soon. You need to sort the Jack pictures from the mountain pictures—both the ones you've taken and the ones you will take—and then store both somewhere other than your phone. But to be clear: there are a LOT of both. You don't have the time to sort them manually. So you decide to hire a programmer to help you.

Here's how it will work. First, we define the task to be completed. In this case, easy enough: sort the Jack pictures from the mountain pictures. Now, start with what is called an *untrained* AI—in this case, a basic program that can sort photos, but in a way that is willy-nilly. The program, in other words, can place your photos into files, but at random. Like a child who can vocalize but not speak, the program is *untrained*. Next step: provide the program with a lot of your pre-sorted and labeled photos. The more, the better. These photos are—to use the technical term—the *training data* for the AI. The fact that your training data is pre-sorted and labeled is crucial—*you* know how the photos should be sorted, and so can use them to train your AI.

Now that we have our task, pre-sorted training data, and untrained AI, we can start training. The process here is in some ways very much like teaching a human child to talk. Initially, the program will sort photos in a way that seems entirely random, like an eighteen-month-old babbling random syllables. But also like a toddler learning to speak, the program gets some things right. So, we provide feedback to the untrained AI—again, we already sorted these photos, so we can "tell" the AI when it sorted correctly, and where it missed the mark. Eventually, through this feedback, the program—like a toddler experimenting with "goo goo," "gaga," "dada," and "mama"—gets better at what you are training it to do. And better. To the point that it has “learned” the task it was assigned. The toddler has learned to speak. Or, in the case of our program, to sort Jack photos from mountain photos.

At this point, the AI is fully *trained*. We can now start feeding it your new photos, photos *outside* the initial training data set you provided. If trained correctly, the program will sort these photos into the appropriate files. Of course, it might get some things wrong. But we can tweak it along the way, and the result will be an increasingly capable program, one “that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.” We’ve built a fully-trained AI. A bit of a dull AI, true. But an AI nonetheless.

There are, however, crucial limitations to the program you have just developed. And these limitations will be crucial to understanding flashier forms of artificial intelligence.

First, the program was developed to do one task, and one task alone: sort Jack pictures from mountain pictures. It does an excellent job at that. But *just* that. What if the photo you give it contains neither Jack nor mountains? Throw a selfie in the mix, and you’ll flummox the AI. It would be like requisitioning a Model T assembly line to produce axes. You won’t like the product since that’s not what it was built to do.

Second, the way the AI works will not be entirely comprehensible, even to the programmer. For someone new to AI, that claim is astounding. Yet it’s true. Here’s why: during the training stage, recall, we were “teaching” the AI to sort photos by tweaking its technique until it got better and better. The hiccup: those “tweaks” were not all completed by a human. Indeed, when training an AI—especially in its most sophisticated forms—the “tweaks” are performed almost entirely by the program itself, in subtle and complex ways. And that makes it near-impossible to track, even by the people building it. Think about it this way: picture yourself in the cockpit of a plane, surrounded by tiny knobs. You are hoping to get the plane to maintain a course of north-northeast. Currently, the plane is off track, so you start turning knobs, clueless as to what they do. To make things more complicated, when you turn one knob, this affects the operation of the other knobs in subtle ways. So when you move Knob 1314 to the right, this affects (in a minor but crucial way) the way Knob 1502 operates. You can't, then, keep track of all the effects of your adjustments. Yet you can turn knobs until you get the plane flying the right direction. So that's what you do. Until the plane maintains a course north-northeast. How? You're clueless. You just know you're heading in the right direction. At this point, your situation is similar to the one that a programmer stands in relation to a fully-trained AI. The programmer can determine *that* the program is doing what it is supposed to do. But how is the AI doing this? The programmer is clueless. And his situation is even worse than yours in the cockpit. After all, *you* were the one turning the knobs in the cockpit. The programmer, by contrast, was letting the AI turn the knobs on its own (often millions or billions of these "parameters"). The result? The way your simple photo-sorting AI works will not be entirely understandable, even to the programmer.

A third limitation: your program is inherently limited by *your* set of pictures. Put technically: your fully-trained AI is inherently limited by its training data. The training data, after all, captures the entire world for the program that gets created from it. That's all the program knows. For your program, the world thus consists in *your* Jack pictures and *your* mountain pictures. Nothing else. Throw some other set of dog and mountain pictures at it, and it *may* sort them well . . . or it may sort them very poorly. Those pictures, after all, lie outside the world of the program.

Yawn, you might be thinking. What's the big deal? Truth be told, these limitations probably *aren't* a big deal. After all, not much depends on the limitations of a simple photo-sorting program. Yet in an era of AI, we will be—and indeed, already are—relying on AI to do a lot more than organize digital scrapbooks. AI has scaled up. Way up. And with it, the challenges have scaled up as well.

Artificial Intelligence at Scale

Compare a program that can sort photos on your phone with the kinds of AI that opened our discussion in this chapter, and that have captured our collective imaginations: a chatbot composing essays on regret; code crafting Cubist cats; algorithms arranging Beatles-flavored anthems. There are some crucial differences between simple AI and the scaled-up versions.

First, the tasks to which flashier platforms are assigned are much more complex. Sorting shots of Jack from pictures of the foothills? Child's play. Crafting Transcendentalist essays or composing paintings in Impressionist style or producing drone shots of Gold Rush-era towns? Definitely not child's play.

Second, many of the headline-grabbing platforms are not merely *sorting* information. Rather, they are *generating* it. They are creating something new. That's what the G in ChatGPT means—*generative*. And as anyone who has tried writing a song or essay can tell you: writing an original is magnitudes more complicated than using an existing one. Generative AI has been with us for a while: for years email apps have been generating sentences as we write them. Yet generative AI has taken large leaps forward recently, largely due to the development of a "deep learning architecture" called a transformer (the T of Chat GPT).[2] While your email app can predict how a sentence might wrap up, a transformer can predict how an entire essay on Jane Austen's view of the good life might unfold.

A third difference between your photo sorting program and a generative AI like ChatGPT: a platform like ChatGPT has been trained on *way* more information than your simple photo sorting program. Sure, you may think you have a lot of pictures on your phone, and that the training data set of your program was substantial. But contrast your phone's storage with ChatGPT's training data set: millions of eBooks, a big chunk of Reddit, all of Wikipedia, and snapshots of the entire internet. Yes, the entire internet, accessed from a non-profit company called Common Crawl, that does its best to capture internet landscapes. That training data set versus your photo album? Like the difference between a file cabinet and a warehouse full of files.

Of course, the creativity, complexity, and training data size of large-scale, generative AI are all related. The fact that ChatGPT *generates* something is a big part of what makes it complex. And when it comes to AI, the more complex function you need it to perform, the larger training data set you need to provide. Suppose, for example, that you wanted your photo-sorting program to stack pictures into *four* piles: Jack shots, mountain vistas, selfies, and pictures of the pizzas you've started baking (yes, your phone's photo album is a bit eccentric—but whose isn't?). You'd need to provide it with more selfies and pizza shots—a lot more—so it could hone on in the features that make these pictures different. More complicated task—more data needed and more metaphorical knobs to adjust the algorithm. Same with a generative AI such as ChatGPT.

Despite these differences, however, generative AI overlaps in significant ways with your simple photo-sorting program. Both are built for a specific purpose: your photo-sorting program, to sort photos. ChatGPT, to craft prose that looks human-generated. Both, in their untrained versions, are capable of little. But both achieve their purposes through exposure to training data sets and a training period. And when fully trained, both can make inferences from an input to a relevant set of outputs. Finally, like simple AI, even the most sophisticated AI systems face important limitations. And they are (more or less) the same limitations faced by the simplest AI.

In the first place, large-scale, generative AI systems are limited to the task for which they were built. Take ChatGPT. ChatGPT was created to generate text that *appears* to have been written by a human being. And it does that remarkably well. It writes essays and poems and recipes that really do *look* like a human being wrote them. But here's what often gets forgotten. Humans use language to do things that go beyond what ChatGPT was trained to do. For example, we use writing to convey truths about the world around us, help guide others in the correct way to complete a task, convey meaning . . . the list goes on. Rely on ChatGPT to do *those* things, and things can go poorly. For example: while you can get ChatGPT to generate something that *looks* like a recipe, there's no guarantee that recipe will taste any good. The platform, after all, was trained to produce text that looks like a recipe, not provide recipes that taste good. Likewise, when teaching, I take care to explain to my students what ChatGPT can and cannot do. In particular, while ChatGPT does a bang-up job of churning out text that *looks* like a college essay, it wasn't trained to do other things that an excellent college essay is supposed to do: convey accurate information; include personal reflection; and demonstrate creative and reflective thinking. It can do the big things. But not many of the little things that add in the human touch. Or take its attempts at poetry: while working on this chapter, I asked ChatGPT to write a short poem about nostalgia for a simpler time. What can I say? There's nothing like writing about the future that can make a guy long for the past. Here's an excerpt of what it spat out:

Yearning for Simpler Days

In the whispers of fading light,
Nostalgia takes its gentle flight,
To a realm where memories gleam,
A simpler time, a cherished dream.

Through hazy lenses of the past,
We yearn for moments that won't last,
For in our hearts, the ember glows,
A longing for the path that goes.

But let us find solace in the truth,
That memories hold eternal youth,
And as we journey on life's tide,
The simpler times forever abide.

So let nostalgia's gentle sway,
Inspire us on this modern day,
To embrace the lessons that we find,
In memories of a simpler time.

First reaction: the poem sure *sounds* like a person wrote it. It's definitely a poem-shaped object. And it even has a few good moments: the idea that "memories hold eternal youth" is comforting, if a bit saccharine. But look closer, and things get iffy. For one thing, almost every line is boiler-plate and cliché. Cheesy, even. The end rhymes are predictable, the kind spun by a third-grader or amateur free-style lyricist. And still other parts don't quite make sense (What about this "path that goes"? Where does it go?). The poem, in short, looks a lot like a poem. Yet it fails at many things poems should do. Why? Because ChatGPT wasn't *trained* to write meaningful, creative, original, or insightful poetry—it was trained to create text that looks like a human wrote it. And that's exactly what it did. To ask it to do anything else would be like asking your picture-sorting program to sort photos of deciduous from coniferous trees, or a Model T assembly line to start producing axes.

A second limitation of generative AI: its operations are opaque, even to those who build them. We have already seen why. Picture the pilot's cockpit again, but this time with tenfold or ten thousandfold the dials of the previous one. Again, turning any one dial can affect the operations of others. Again, the dials are turned more or less at random, and judged by the results. Also again: it isn't some *human* who is doing that fine-tuning. No, remember that all of this is done automatically, by the program itself during the training phase. What this means: when the knobs get adjusted just right, it is impossible for the creator to discover precisely what did the trick. The programmer can simply shrug and say: well, apparently, somewhere in there, something got adjusted right! The lesson: the more sophisticated the AI, the more clueless its creator is about what ensures its success. So, once again—but to an even larger degree—while we can recognize *when* an AI is performing the task we want it to do, it is near impossible to determine *how* it is doing it or what the unintended consequences might be.

One final limitation of AI: even the most sophisticated AI systems are limited to the world of their training data sets. Just as your photo-sorting program would be clueless when confronted with a photo of a cat or a Ferrari, so too a platform like ChatGPT is clueless in the face of anything outside its training data set. True, any generative AI like ChatGPT has a large training data set, and so draws upon a large world. But even *its* world has limits. ChatGPT, for example, was trained largely on text from the internet. And, it turns out (said in a dry tone), the internet is not the entire world.

From AI to Ethics

We live in a new era: the era of artificial intelligence. In any new era, we are tempted to charge forward or retreat to a more comfortable past. Both strategies are too easy. A better way forward: understand the technology on its own terms and interrogate the questions it raises. We have worked on the first task in this chapter. We turn to the second task in the next. AI, we'll see, challenges our humanity along several fronts. Understanding the contours of these challenges is a crucial first step to mounting a response. In doing so, we will inevitably learn something about ourselves along the way. And come closer to achieving our aim: staying human in an era of AI.