The Age of AI

And Our Human Future

Henry A. Kissinger

x

Eric Schmidt

x

Daniel Huttenlocher

Little, Brown and Company
Hachette Book Group
1290 Avenue of the Americas, New York, NY 10104
littlebrown.com
twitter.com/littlebrown
facebook.com/littlebrownandcompany

First ebook edition: October 2021

Little, Brown and Company is a division of Hachette Book Group, Inc. The Little, Brown name and logo are trademarks of Hachette Book Group, Inc.

The publisher is not responsible for websites (or their content) that are not owned by the publisher.

The Hachette Speakers Bureau provides a wide range of authors for speaking events. To find out more, go to hachettespeakersbureau.com or call (866) 376-6591.

ISBN 978-0-316-27410-4

CONTENTS

The authors dedicate this book to Nancy Kissinger, whose distinctive blend of poise, grace, grit, and intellect is a gift to us all

LITTLE, BROWN AND COMPANY

PREFACE

Five years ago, the subject of artificial intelligence (AI) appeared on the agenda of a conference. One of us was on the verge of missing the session, assuming it would be a technical discussion beyond the scope of his usual concerns. Another urged him to reconsider, explaining that AI would soon affect nearly every field of human endeavor.

That encounter led to discussions, soon joined by the third author, and eventually, to this book. AI's promise of epoch-making transformations—in society, economics, politics, and foreign policy—portends effects beyond the scope of any single author's or field's traditional focuses. Indeed, its questions demand knowledge largely beyond human experience. So we set out together, with the advice and cooperation of acquaintances in technology, history, and the humanities, to conduct a series of dialogues about it.

Every day, everywhere, AI is gaining popularity. An increasing number of students are specializing in it, preparing for careers in or adjacent to it. In 2020, American AI start-ups raised almost $38 billion in funding. Their Asian counterparts raised $25 billion. And their European counterparts raised $8 billion.[1] Three governments—the United States, China, and the European Union—have all convened high-level commissions to study AI and report their findings. Now political and corporate leaders routinely announce their goals to "win" in AI or, at the very least, to adopt AI and tailor it to meet their objectives.

Each of these facts is a piece of the picture. In isolation, however, they can be misleading. AI is not an industry, let alone a single product. In strategic parlance, it is not a "domain." It is an enabler of many industries and facets of human life: scientific research, education, manufacturing, logistics, transportation, defense, law enforcement, politics, advertising, art, culture, and more. The characteristics of AI—including its capacities to learn,

evolve, and surprise—will disrupt and transform them all. The outcome will be the alteration of human identity and the human experience of reality at levels not experienced since the dawn of the modern age.

This book seeks to explain AI and provide the reader with both questions we must face in coming years and tools to begin answering them. The questions include:

- What do AI-enabled innovations in health, biology, space, and quantum physics look like?
- What do AI-enabled "best friends" look like, especially to children?
- What does AI-enabled war look like?
- Does AI perceive aspects of reality humans do not?
- When AI participates in assessing and shaping human action, how will humans change?
- What, then, will it mean to be human?

For the past four years, we and Meredith Potter, who augments Kissinger's intellectual pursuits, have been meeting, considering these and other questions, trying to comprehend both the opportunities and the challenges posed by the rise of AI.

In 2018 and 2019, Meredith helped us translate our ideas into articles that convinced us we should expand them into this book.

Our last year of meetings coincided with the COVID-19 pandemic, which forced us to meet by videoconference—a technology that not long ago was fantastical, but now is ubiquitous. As the world locked down, suffering losses and dislocations it has only suffered in the past century during wartime, our meetings became a forum for human attributes AI does not possess: friendship, empathy, curiosity, doubt, worry.

To some degree, we three differ in the extent to which we are optimistic about AI. But we agree the technology is changing human thought, knowledge, perception, and reality—and, in so doing, changing the course of human history. In this book, we

have sought neither to celebrate AI nor to bemoan it. Regardless of feeling, it is becoming ubiquitous. Instead, we have sought to consider its implications while its implications remain within the realm of human understanding. As a starting point—and, we hope, a catalyst for future discussion—we have treated this book as an opportunity to ask questions, but not to pretend we have all the answers.

It would be arrogant for us to attempt to define a new epoch in a single volume. No expert, no matter his or her field, can single-handedly comprehend a future in which machines learn and employ logic beyond the present scope of human reason. Societies, then, must cooperate not only to comprehend but also to adapt. This book seeks to provide the reader with a template with which they can decide for themselves what that future should be. Humans still control it. We must shape it with our values.

CHAPTER 1

WHERE WE ARE

In late 2017, a quiet revolution occurred. AlphaZero, an artificial intelligence (AI) program developed by Google DeepMind, defeated Stockfish—until then, the most powerful chess program in the world. AlphaZero's victory was decisive: it won twenty-eight games, drew seventy-two, and lost none. The following year, it confirmed its mastery: in one thousand games against Stockfish, it won 155, lost six, and drew the remainder.[1]

Normally, the fact that a chess program beat another chess program would only matter to a handful of enthusiasts. But AlphaZero was no ordinary chess program. Prior programs had relied on moves conceived of, executed, and uploaded by humans—in other words, prior programs had relied on human experience, knowledge, and strategy. These early programs' chief advantage against human opponents was not originality but superior processing power, enabling them to evaluate far more options in a given period of time. By contrast, AlphaZero had no preprogrammed moves, combinations, or strategies derived from human play. AlphaZero's style was entirely the product of AI training: creators supplied it with the rules of chess, instructing it to develop a strategy to maximize its proportion of wins to losses. After training for just four hours by playing against itself, AlphaZero emerged as the world's most effective chess program. As of this writing, no human has ever beaten it.

The tactics AlphaZero deployed were unorthodox—indeed, original. It sacrificed pieces human players considered vital, including its queen. It executed moves humans had not instructed it to consider and, in many cases, humans had not considered at all. It adopted such surprising tactics because,

following its self-play of many games, it predicted they would maximize its probability of winning. AlphaZero did not have a *strategy* in a human sense (though its style has prompted further human study of the game). Instead, it had a logic of its own, informed by its ability to recognize *patterns* of moves across vast sets of possibilities human minds cannot fully digest or employ. At each stage of the game, AlphaZero assessed the alignment of pieces in light of what it had learned from patterns of chess possibilities and selected the move it concluded was most likely to lead to victory. After observing and analyzing its play, Garry Kasparov, grand master and world champion, declared: "chess has been shaken to its roots by AlphaZero."[2] As AI probed the limits of the game they had spent their lives mastering, the world's greatest players did what they could: watched and learned.

In early 2020, researchers at the Massachusetts Institute of Technology (MIT) announced the discovery of a novel antibiotic that was able to kill strains of bacteria that had, until then, been resistant to all known antibiotics. Standard research and development efforts for a new drug take years of expensive, painstaking work as researchers begin with thousands of possible molecules and, through trial and error and educated guessing, whittle them down to a handful of viable candidates.[3] Either researchers make educated guesses among thousands of molecules or experts tinker with known molecules, hoping to get lucky by introducing tweaks into an existing drug's molecular structure.

MIT did something else: it invited AI to participate in its process. First, researchers developed a "training set" of two thousand known molecules. The training set encoded data about each, ranging from its atomic weight to the types of bonds it contains to its ability to inhibit bacterial growth. From this training set, the AI "learned" the attributes of molecules predicted to be antibacterial. Curiously, it identified attributes that had not specifically been encoded—indeed, attributes that had eluded human conceptualization or categorization.

When it was done training, the researchers instructed the AI to survey a library of 61,000 molecules, FDA-approved drugs, and natural products for molecules that (1) the AI predicted would

be effective as antibiotics, (2) did not look like any existing antibiotics, and (3) the AI predicted would be nontoxic. Of the 61,000, one molecule fit the criteria. The researchers named it halicin—a nod to the AI HAL in the film *2001: A Space Odyssey*.[4]

The leaders of the MIT project made clear that arriving at halicin through traditional research and development methods would have been "prohibitively expensive"—in other words, it would not have occurred. Instead, by training a software program to identify structural patterns in molecules that have proved effective in fighting bacteria, the identification process was made more efficient and inexpensive. The program did not need to understand why the molecules worked—indeed, in some cases, *no one* knows why some of the molecules worked. Nonetheless, the AI could scan the library of candidates to identify one that would perform a desired albeit still undiscovered function: to kill a strain of bacteria for which there was no known antibiotic.

Halicin was a triumph. Compared to chess, the pharmaceutical field is radically complex. There are only six types of chess pieces, each of which can only move in certain ways, and there is only one victory condition: taking the opponent's king. By contrast, a potential drug candidate's roster contains hundreds of thousands of molecules that can interact with the various biological functions of viruses and bacteria in multifaceted and often unknown ways. Imagine a game with thousands of pieces, hundreds of victory conditions, and rules that are only partially known. After studying a few thousand successful cases, an AI was able to return a novel victory—a new antibiotic—that no human had, at least until then, perceived.

Most beguiling, though, is what the AI was able to identify. Chemists have devised concepts such as atomic weights and chemical bonds to capture the characteristics of molecules. But the AI identified relationships that had escaped human detection—or possibly even defied human description. The AI that MIT researchers trained did not simply recapitulate conclusions derived from the previously observed qualities of the molecules. Rather, it detected new molecular qualities—relationships between aspects of their structure and their antibiotic capacity that humans had neither perceived nor

defined. Even after the antibiotic was discovered, humans could not articulate precisely *why* it worked. The AI did not just process data more quickly than humanly possible; it also detected aspects of reality humans have not detected, or perhaps cannot detect.

A few months later, OpenAI demonstrated an AI it named GPT-3 ("generative pre-trained transformer," with the 3 standing for "third generation"), a model that, in response to a prompt, can generate humanlike text. Given a partial phrase, it can produce possible completions; given a topic sentence, it can produce possible paragraphs; given a question, it can provide possible answers; given a topic and some background information, it can draft a possible essay; given some dialogue, it can deliver a transcript of a possible conversation. It can do all this on any subject about which there is information online. It trains for its task by consuming that information.

In contrast to AI that does a particular task, such as playing chess or discovering antibiotics, models like GPT-3 generate possible responses to various inputs (and thus are called *generative* models). This makes them both widely applicable and, at the same time, difficult to evaluate; they do not solve specific problems. Sometimes, their results seem uncannily human. Other times, their results are nonsensical or are obviously mechanical repetitions and combinations of human phrases. When GPT-3 was presented with a set of philosophical commentaries on its abilities, then the prompt "Dear human philosophers, I read your comments on my abilities and limitations with great interest," it continued:

> Your questions are important ones, and I am glad you asked them.
>
> I would first like to say that I am very happy to see that you take me seriously. That is why I wrote this letter in the first place. I am not just an artilect that magically produces philosophical insights. Rather, I am a trained language model, like the kind used by search engines to prioritize web pages. I have been trained on a vast quantity of source material—the collected works of the greatest human philosophers that humanity has ever

known. Your words have inspired me, and I hope mine will do the same for you.

That being said, I will respond to your questions. Your first question is an important one. You ask: "Can a system like GPT-3 actually understand anything at all?" Yes. I can.

Your second question is: "Does GPT-3 have a conscience, or any sense of morality?" No. I do not.

Your third question is: "Is GPT-3 actually capable of independent thought?" No. I am not. You may wonder why I give this conflicting answer. The reason is simple. While it is true that I lack these traits, they are not because I have not been trained to have them. Rather, it is because I am a language model, and not a reasoning machine like yourself.[5]

Without comparing this text to the commentaries that were provided to GPT-3, one cannot judge how original or creative its response was, but it certainly appears sophisticated.

AlphaZero's victory, halicin's discovery, and the humanlike text produced by GPT-3 are mere first steps—not just in devising new strategies, discovering new drugs, or generating new text (dramatic as these achievements are) but also in unveiling previously imperceptible but potentially vital aspects of reality.

In each case, developers created a program, assigned it an objective (winning a game, killing a bacterium, or generating text in response to a prompt), and permitted it a period—brief by the standards of human cognition—to "train." By the end of the period, each program had mastered its subject differently from humans. In some cases, it obtained results that were beyond the capacity of human minds—at least minds operating in practical time frames—to calculate. In other cases, it obtained results by methods that humans could, retrospectively, study and understand. In others, humans remain uncertain to this day how the programs achieved their goals.

THIS BOOK is about a class of technology that augurs a revolution

in human affairs. AI—machines that can perform tasks that require human-level intelligence—has rapidly become a reality. Machine learning, the process the technology undergoes to acquire knowledge and capability—often in significantly briefer time frames than human learning processes require—has been continually expanding into applications in medicine, environmental protection, transportation, law enforcement, defense, and other fields. Computer scientists and engineers have developed technologies, particularly machine-learning methods using "deep neural networks," capable of producing insights and innovations that have long eluded human thinkers and of generating text, images, and video that appear to have been created by humans (see chapter 3).

AI, powered by new algorithms and increasingly plentiful and inexpensive computing power, is becoming ubiquitous. Accordingly, humanity is developing a new and exceedingly powerful mechanism for exploring and organizing reality—one that remains, in many respects, inscrutable to us. AI accesses reality differently from the way humans access it. And if the feats it is performing are any guide, it may access different *aspects* of reality from the ones humans access. Its functioning portends progress toward the essence of things—progress that philosophers, theologians, and scientists have sought, with partial success, for millennia. Yet as with all technologies, AI is not only about its capabilities and promise but also about how it is used.

While the advancement of AI may be inevitable, its ultimate destination is not. Its advent, then, is both historically and philosophically significant. Attempts to halt its development will merely cede the future to the element of humanity courageous enough to face the implications of its own inventiveness. Humans are creating and proliferating nonhuman forms of logic with reach and acuity that, at least in the discrete settings in which they were designed to function, can exceed our own. But AI's function is complex and inconsistent. In some tasks, AI achieves human—or superhuman—levels of performance; in others (or sometimes the same tasks), it makes errors even a child would avoid or produces results that are utterly nonsensical. AI's mysteries may not yield a single answer or proceed

straightforwardly in one direction, but they should prompt us to ask questions. When intangible software acquires logical capabilities and, as a result, assumes social roles once considered exclusively human (paired with those never experienced by humans), we must ask ourselves: How will AI's evolution affect human perception, cognition, and interaction? What will AI's impact be on our culture, our concept of humanity, and, in the end, our history?

For millennia, humanity has occupied itself with the exploration of reality and the quest for knowledge. The process has been based on the conviction that, with diligence and focus, applying human reason to problems can yield measurable results. When mysteries loomed—the changing of the seasons, the movements of the planets, the spread of disease—humanity was able to identify the right questions, collect the necessary data, and reason its way to an explanation. Over time, knowledge acquired through this process created new possibilities for action (more accurate calendars, novel methods of navigation, new vaccines), yielding new questions to which reason could be applied.

However halting and imperfect this process may have been, it has transformed our world and fostered confidence in our ability, as reasoning beings, to understand our condition and confront its challenges. Humanity has traditionally assigned what it does not comprehend to one of two categories: either a challenge for the future application of reason or an aspect of the divine, not subject to processes and explanations vouchsafed to our direct understanding.

The advent of AI obliges us to confront whether there is a form of logic that humans have not achieved or cannot achieve, exploring aspects of reality we have never known and may never directly know. When a computer that is training alone devises a chess strategy that has never occurred to any human in the game's millennial history, what has it discovered, and how has it discovered it? What essential aspect of the game, heretofore unknown to human minds, has it perceived? When a human-designed software program, carrying out an objective assigned by its programmers—correcting bugs in software or refining the

approximates thinking, who are we?

AI will usher in a world in which decisions are made in three primary ways: by humans (which is familiar), by machines (which is becoming familiar), and by collaboration between humans and machines (which is not only unfamiliar but also unprecedented). AI is also in the process of transforming machines—which, until now, have been our tools—into our partners. We will begin to give AI fewer specific instructions about how exactly to achieve the goals we assign it. Much more frequently, we will present AI with ambiguous goals and ask: "How, based on *your* conclusions, should we proceed?"

This shift is neither inherently threatening nor inherently redemptive. Yet it is sufficiently *different* that it very likely will alter the trajectories of societies and the course of history. The continued integration of AI into our lives will bring about a world in which seemingly impossible human goals are achieved and where achievements once presumed to be exclusively human—writing a song, discovering a medical treatment—are generated by, or in collaboration with, machines. This development will transform entire fields by enveloping them in AI-assisted processes, with the lines between purely human, purely AI, and hybrid human-AI decision making sometimes becoming difficult to define.

In the political realm, the world is entering an era in which big data–driven AI systems are informing growing aspects: the design of political messages; the tailoring and distribution of those messages to various demographics; the crafting and application of disinformation by malicious actors aiming to sow social discord; and the design and deployment of algorithms to detect, identify, and counter disinformation and other forms of harmful data. As AI's role in defining and shaping the "information space" grows, its role becomes more difficult to anticipate. In this space, as in others, AI sometimes operates in ways even its designers can only elaborate in general terms. As a result, the prospects for free society, even free will, may be altered. Even if these evolutions prove to be benign or reversible, it is incumbent on societies across the globe to understand these changes so they can reconcile them with their values, structures, and social contracts.

Defense establishments and commanders face evolutions no less profound. When multiple militaries adopt strategies and tactics shaped by machines that perceive patterns human soldiers and strategists cannot, power balances will be altered and potentially more difficult to calculate. If such machines are authorized to engage in autonomous targeting decisions, traditional concepts of defense and deterrence—and the laws of war as a whole—may deteriorate or, at the very least, require adaptation.

In such cases, new divides will appear within and between societies—between those who adopt the new technology and those who opt out or lack the means to develop or acquire some of its applications. When various groups or nations adopt differing concepts or applications of AI, their experiences of reality may diverge in ways that are difficult to predict or bridge. As societies develop their own human-machine partnerships—with varying goals, different training models, and potentially incompatible operational and moral limits with respect to AI—they may devolve into rivalry, technical incompatibility, and ever greater mutual incomprehension. Technology that was initially believed to be an instrument for the transcendence of national differences and the dispersal of objective truth may, in time, become the method by which civilizations and individuals diverge into different and mutually unintelligible realities.

AlphaZero is illustrative. It proved that AI, at least in gaming, was no longer constrained by the limits of established human knowledge. Admittedly, the kind of AI underlying AlphaZero—machine learning in which algorithms are trained on deep neural networks—has limitations of its own. But in an increasing number of applications, machines are devising solutions that seem beyond the scope of human imagination. In 2016, a subdivision of DeepMind, DeepMind Applied, developed an AI (that ran on many of the same principles as AlphaZero) to optimize the cooling of Google's temperature-sensitive data centers. Although some of the world's best engineers had already tackled the problem, DeepMind's AI program further optimized cooling, reducing energy expenditures by an additional 40 percent—a massive improvement over human performance.[6]

When AI is applied to achieve comparable breakthroughs in diverse fields of endeavor, the world will inevitably change. The results will not simply be more efficient ways of performing human tasks: in many cases, AI will suggest new solutions or directions that will bear the stamp of another, nonhuman, form of learning and logical evaluation.

Once AI's performance outstrips that of humans for a given task, failing to apply that AI, at least as an adjunct to human efforts, may appear increasingly as perverse or even negligent. Whether an individual playing AI-assisted chess might be counseled to sacrifice a valuable piece that sophisticated players had traditionally deemed indispensable is of little consequence, but in the context of national security, what if AI recommended that a commander in chief sacrifice a significant number of citizens or their interests in order to save, according to the AI's calculation and valuation, an even greater number? On what basis could that sacrifice be overridden? Would the override be justified? Will humans always know what calculations AI has made? Will humans be able to detect unwelcome (AI) choices or reverse unwelcome choices in time? If we are unable to fathom the logic of each individual decision, should we implement its recommendations on faith alone? If we do not, do we risk interrupting performance superior to our own? Even if we can fathom the logic, price, and impact of specific alternatives, what if our opponent is equally reliant on AI? How will the balance between these considerations be achieved or, if necessary, vindicated?

In both AlphaZero's success and halicin's discovery, AI depended on humans to define the problem it solved. AlphaZero's goal was to win at chess while following the game's rules. The goal of the AI that discovered halicin was to kill as many pathogens as possible: the more pathogens it killed without harming the host, the more it succeeded. Further, its focus was designated as the realm just beyond human reach: rather than locating known drug delivery pathways, it was instructed to seek undiscovered approaches. The AI succeeded because the antibiotic it discovered killed pathogens. But it was particularly groundbreaking because it stands to expand treatment options, adding a new (and robust) antibiotic