How the Enlightenment Ends

Philosophically, intellectually—in every way—human society is unprepared for the rise of artificial intelligence.

By Henry A. Kissinger - The Atlantic, June 2018 Issue

Three years ago, at a conference on transatlantic issues, the subject of artificial intelligence appeared on the agenda. I was on the verge of skipping that session—it lay outside my usual concerns—but the beginning of the presentation held me in my seat.

The speaker described the workings of a computer program that would soon challenge international champions in the game Go. I was amazed that a computer could master Go, which is more complex than chess. In it, each player deploys 180 or 181 pieces (depending on which color he or she chooses), placed alternately on an initially empty board; victory goes to the side that, by making better strategic decisions, immobilizes his or her opponent by more effectively controlling territory.

The speaker insisted that this ability could not be preprogrammed. His machine, he said, learned to master Go by training itself through practice. Given Go's basic rules, the computer played innumerable games against itself, learning from its mistakes and refining its algorithms accordingly. In the process, it exceeded the skills of its human mentors. And indeed, in the months following the speech, an AI program named AlphaGo would decisively defeat the world's greatest Go players.

As I listened to the speaker celebrate this technical progress, my experience as a historian and occasional practicing statesman gave me pause. What would be the impact on history of self-learning machines—machines that acquired knowledge by processes particular to themselves, and applied that knowledge to ends for which there may be no category of human understanding? Would these machines learn to communicate with one another? How would choices be made among emerging options? Was it possible that human history might go the way of the Incas, faced with a Spanish culture incomprehensible and even awe-inspiring to them? Were we at the edge of a new phase of human history?

Aware of my lack of technical competence in this field, I organized a number of informal dialogues on the subject, with the advice and cooperation of acquaintances in technology and the humanities. These discussions have caused my concerns to grow.

Heretofore, the technological advance that most altered the course of modern history was the invention of the printing press in the 15th century, which allowed the search for empirical knowledge to supplant liturgical doctrine, and the Age of Reason to gradually supersede the Age of Religion. Individual insight and scientific knowledge replaced faith as the principal criterion of human consciousness. Information was stored and systematized in expanding libraries. The Age of Reason originated the thoughts and actions that shaped the contemporary world order.

But that order is now in upheaval amid a new, even more sweeping technological revolution whose consequences we have failed to fully reckon with, and whose culmination may be a world relying on machines powered by data and algorithms and ungoverned by ethical or philosophical norms.

The internet age in which we already live prefigures some of the questions and issues that AI will only make more acute. The Enlightenment sought to submit traditional verities to a liberated, analytic human reason. The internet's purpose is to ratify knowledge through the accumulation and manipulation of ever expanding data. Human cognition loses its personal character. Individuals turn into data, and data become regnant.

Users of the internet emphasize retrieving and manipulating information over contextualizing or conceptualizing its meaning. They rarely interrogate history or philosophy; as a rule, they demand information relevant to their immediate practical needs. In the process, search-engine algorithms acquire the capacity to predict the preferences of individual clients, enabling the algorithms to personalize results and make them available to other parties for political or commercial purposes. Truth becomes relative. Information threatens to overwhelm wisdom.

Inundated via social media with the opinions of multitudes, users are diverted from introspection; in truth many technophiles use the internet to avoid the solitude they dread. All of these pressures weaken the fortitude required to develop and sustain convictions that can be implemented only by traveling a lonely road, which is the essence of creativity.

The impact of internet technology on politics is particularly pronounced. The ability to target micro-groups has broken up the previous consensus on priorities by permitting a focus on specialized purposes or grievances. Political leaders, overwhelmed by niche pressures, are deprived of time to think or reflect on context, contracting the space available for them to develop vision.

The digital world's emphasis on speed inhibits reflection; its incentive empowers the radical over the thoughtful; its values are shaped by subgroup consensus, not by introspection. For all its achievements, it runs the risk of turning on itself as its impositions overwhelm its conveniences.

As the internet and increased computing power have facilitated the accumulation and analysis of vast data, unprecedented vistas for human understanding have emerged. Perhaps most significant is the project of producing artificial intelligence—a technology capable of inventing and solving complex, seemingly abstract problems by processes that seem to replicate those of the human mind.

This goes far beyond automation as we have known it. Automation deals with means; it achieves prescribed objectives by rationalizing or mechanizing instruments for reaching them. AI, by contrast, deals with ends; it establishes its own objectives. To the extent that its achievements are in part shaped by itself, AI is inherently unstable. AI systems, through their very operations, are in constant flux as they acquire and instantly analyze new data, then seek to improve themselves on the basis of that analysis. Through this process,

artificial intelligence develops an ability previously thought to be reserved for human beings. It makes strategic judgments about the future, some based on data received as code (for example, the rules of a game), and some based on data it gathers itself (for example, by playing 1 million iterations of a game).

The driverless car illustrates the difference between the actions of traditional human-controlled, software-powered computers and the universe AI seeks to navigate. Driving a car requires judgments in multiple situations impossible to anticipate and hence to program in advance. What would happen, to use a well-known hypothetical example, if such a car were obliged by circumstance to choose between killing a grandparent and killing a child? Whom would it choose? Why? Which factors among its options would it attempt to optimize? And could it explain its rationale? Challenged, its truthful answer would likely be, were it able to communicate: "I don't know (because I am following mathematical, not human, principles)," or "You would not understand (because I have been trained to act in a certain way but not to explain it)." Yet driverless cars are likely to be prevalent on roads within a decade.

We must expect AI to make mistakes faster—and of greater magnitude—than humans do.

Heretofore confined to specific fields of activity, AI research now seeks to bring about a "generally intelligent" AI capable of executing tasks in multiple fields. A growing percentage of human activity will, within a measurable time period, be driven by AI algorithms. But these algorithms, being mathematical interpretations of observed data, do not explain the underlying reality that produces them. Paradoxically, as the world becomes more transparent, it will also become increasingly mysterious. What will distinguish that new world from the one we have known? How will we live in it? How will we manage AI, improve it, or at the very least prevent it from doing harm, culminating in the most ominous concern: that AI, by mastering certain competencies more rapidly and definitively than humans, could over time diminish human competence and the human condition itself as it turns it into data.

Artificial intelligence will in time bring extraordinary benefits to medical science, clean-energy provision, environmental issues, and many other areas. But precisely because AI makes judgments regarding an evolving, as-yet-undetermined future, uncertainty and ambiguity are inherent in its results. There are three areas of special concern:

First, that AI may achieve unintended results. Science fiction has imagined scenarios of AI turning on its creators. More likely is the danger that AI will misinterpret human instructions due to its inherent lack of context. A famous recent example was the AI chatbot called Tay, designed to generate friendly conversation in the language patterns of a 19-year-old girl. But the machine proved unable to define the imperatives of "friendly" and "reasonable" language installed by its instructors and instead became racist, sexist, and otherwise inflammatory in its responses. Some in the technology world claim that the experiment was ill-conceived and poorly executed, but it illustrates an underlying ambiguity: To what extent is it possible to enable AI to comprehend the context that informs its instructions? What medium could have helped Tay define for itself offensive, a word upon whose meaning

humans do not universally agree? Can we, at an early stage, detect and correct an AI program that is acting outside our framework of expectation? Or will AI, left to its own devices, inevitably develop slight deviations that could, over time, cascade into catastrophic departures?

Second, that in achieving intended goals, AI may change human thought processes and human values. AlphaGo defeated the world Go champions by making strategically unprecedented moves—moves that humans had not conceived and have not yet successfully learned to overcome. Are these moves beyond the capacity of the human brain? Or could humans learn them now that they have been demonstrated by a new master?

Before AI began to play Go, the game had varied, layered purposes: A player sought not only to win, but also to learn new strategies potentially applicable to other of life's dimensions. For its part, by contrast, AI knows only one purpose: to win. It "learns" not conceptually but mathematically, by marginal adjustments to its algorithms. So in learning to win Go by playing it differently than humans do, AI has changed both the game's nature and its impact. Does this single-minded insistence on prevailing characterize all AI?

Other AI projects work on modifying human thought by developing devices capable of generating a range of answers to human queries. Beyond factual questions ("What is the temperature outside?"), questions about the nature of reality or the meaning of life raise deeper issues. Do we want children to learn values through discourse with untethered algorithms? Should we protect privacy by restricting AI's learning about its questioners? If so, how do we accomplish these goals?

If AI learns exponentially faster than humans, we must expect it to accelerate, also exponentially, the trial-and-error process by which human decisions are generally made: to make mistakes faster and of greater magnitude than humans do. It may be impossible to temper those mistakes, as researchers in AI often suggest, by including in a program caveats requiring "ethical" or "reasonable" outcomes. Entire academic disciplines have arisen out of humanity's inability to agree upon how to define these terms. Should AI therefore become their arbiter?

Third, that AI may reach intended goals, but be unable to explain the rationale for its conclusions. In certain fields—pattern recognition, big-data analysis, gaming—AI's capacities already may exceed those of humans. If its computational power continues to compound rapidly, AI may soon be able to optimize situations in ways that are at least marginally different, and probably significantly different, from how humans would optimize them. But at that point, will AI be able to explain, in a way that humans can understand, why its actions are optimal? Or will AI's decision making surpass the explanatory powers of human language and reason? Through all human history, civilizations have created ways to explain the world around them—in the Middle Ages, religion; in the Enlightenment, reason; in the 19th century, history; in the 20th century, ideology. The most difficult yet important question about the world into which we are headed is this: What will become of human

consciousness if its own explanatory power is surpassed by AI, and societies are no longer able to interpret the world they inhabit in terms that are meaningful to them?

How is consciousness to be defined in a world of machines that reduce human experience to mathematical data, interpreted by their own memories? Who is responsible for the actions of AI? How should liability be determined for their mistakes? Can a legal system designed by humans keep pace with activities produced by an AI capable of outthinking and potentially outmaneuvering them?

Ultimately, the term artificial intelligence may be a misnomer. To be sure, these machines can solve complex, seemingly abstract problems that had previously yielded only to human cognition. But what they do uniquely is not thinking as heretofore conceived and experienced. Rather, it is unprecedented memorization and computation. Because of its inherent superiority in these fields, AI is likely to win any game assigned to it. But for our purposes as humans, the games are not only about winning; they are about thinking. By treating a mathematical process as if it were a thought process, and either trying to mimic that process ourselves or merely accepting the results, we are in danger of losing the capacity that has been the essence of human cognition.

The implications of this evolution are shown by a recently designed program, AlphaZero, which plays chess at a level superior to chess masters and in a style not previously seen in chess history. On its own, in just a few hours of self-play, it achieved a level of skill that took human beings 1,500 years to attain. Only the basic rules of the game were provided to AlphaZero. Neither human beings nor human-generated data were part of its process of self-learning. If AlphaZero was able to achieve this mastery so rapidly, where will AI be in five years? What will be the impact on human cognition generally? What is the role of ethics in this process, which consists in essence of the acceleration of choices?

Typically, these questions are left to technologists and to the intelligentsia of related scientific fields. Philosophers and others in the field of the humanities who helped shape previous concepts of world order tend to be disadvantaged, lacking knowledge of AI's mechanisms or being overawed by its capacities. In contrast, the scientific world is impelled to explore the technical possibilities of its achievements, and the technological world is preoccupied with commercial vistas of fabulous scale. The incentive of both these worlds is to push the limits of discoveries rather than to comprehend them. And governance, insofar as it deals with the subject, is more likely to investigate AI's applications for security and intelligence than to explore the transformation of the human condition that it has begun to produce.

The Enlightenment started with essentially philosophical insights spread by a new technology. Our period is moving in the opposite direction. It has generated a potentially dominating technology in search of a guiding philosophy. Other countries have made AI a major national project. The United States has not yet, as a nation, systematically explored its full scope, studied its implications, or begun the process of ultimate learning. This should be given a high national priority, above all, from the point of view of relating AI to humanistic traditions.

AI developers, as inexperienced in politics and philosophy as I am in technology, should ask themselves some of the questions I have raised here in order to build answers into their engineering efforts. The U.S. government should consider a presidential commission of eminent thinkers to help develop a national vision. This much is certain: If we do not start this effort soon, before long we shall discover that we started too late.

Henry A. Kissinger served as national security adviser and secretary of state to Presidents Richard Nixon and Gerald Ford.