The Metamorphosis

theatlantic.com/magazine/archive/2019/08/henry-kissinger-the-metamorphosis-ai/592771

July 11, 2019

Geoffroy de Crécy

Humanity is at the edge of a revolution driven by artificial intelligence. It has the potential to be one of the most significant and far-reaching revolutions in history, yet it has developed out of disparate efforts to solve specific practical problems rather than a comprehensive plan. Ironically, the ultimate effect of this case-by-case problem solving may be the transformation of human reasoning and decision making.

This revolution is unstoppable. Attempts to halt it would cede the future to that element of humanity more courageous in facing the implications of its own inventiveness. Instead, we should accept that AI is bound to become increasingly sophisticated and ubiquitous, and ask ourselves: How will its evolution affect human perception, cognition, and interaction? What will be its impact on our culture and, in the end, our history?

Such questions brought together the three authors of this article: a historian and sometime policy maker; a former chief executive of a major technology company; and the dean of a principal technology-oriented academic institution. We have been meeting for three years to try to understand these issues and their associated riddles. Each of us is convinced of our inability, within the confines of our respective fields of expertise, to fully analyze a future in which machines help guide their own evolution, improving themselves to better solve the problems for which they were designed. So as a starting point—and, we hope, a springboard for wider discussion—we are engaged in framing a more detailed set of questions about the significance of AI's development for human civilization.

The AlphaZero Paradox

Last December, the developers of AlphaZero published their explanation of the process by which the program mastered chess—a process, it turns out, that ignored human chess strategies developed over centuries and classic games from the past. Having been taught the rules of the game,

AlphaZero trained itself entirely by self-play and, in less than 24 hours, became the best chess player in the world—better than grand masters and, until then, the most sophisticated chess-playing computer program in the world. It did so by playing like neither a grand master nor a preexisting program. It conceived and executed moves that both humans and human-trained machines found counterintuitive, if not simply wrong. The founder of the company that created AlphaZero called its performance "chess from another dimension" and proof that sophisticated AI "is no longer constrained by the limits of human knowledge."

Now established chess experts are studying AlphaZero's moves, hoping to incorporate its knowledge into their own play. These studies are practical, but larger philosophical questions also emerge. Among those that are currently unanswerable: How can we explain AlphaZero's capacity to invent a new approach to chess on the basis of a very brief learning period? What was the reality it explored? Will AI lead to an as-yet-unimaginable expansion of familiar reality?

We can expect comparable discoveries by AI in other fields. Some will upend conventional wisdom and standard practices; others will merely tweak them. Nearly all will leave us struggling to understand. Consider the conduct of driverless cars stopped at a traffic light. When cars driven by people inch forward to try to beat the traffic, some driverless cars occasionally join them, though nothing in the rules of driving given to them suggests that they should do so. If this inching-forward has been learned, how and for what purpose? How is it different from what people are taught and learn about waiting at traffic lights? What else might AI learn that it is not "telling" us (because AI does not or cannot explain)? By enabling a process of self-learning for inanimate objects, we do not yet know what we are starting, but we need to find out.

The Nature of the Revolution

Heretofore, digital evolution has relied on human beings to create the software and analyze the data that are so profoundly affecting our lives. Recent advances have recast this process. AI has made it possible to automate an extraordinary range of tasks, and has done so by enabling machines to play a role—an increasingly decisive role—in drawing conclusions from data and then taking action. AI draws lessons from its own experience, unlike traditional software, which can only support human reasoning. The growing transfer of judgment from human beings to machines denotes the revolutionary aspect of AI, as described last year in these pages ("How the Enlightenment Ends," June 2018).

That said, the word *intelligence* does not adequately explain what is occurring, and ascribing anthropomorphic qualities to AI is out of order. AI is neither malicious nor kind; it does not have independently developed intent or goals; it does not engage in self-reflection. What AI can do is to perform well-specified tasks to help discover associations between data and actions, providing solutions for quandaries people find difficult and perhaps impossible. This process creates new forms of automation and in time might yield entirely new ways of thinking.

Yet AI systems today, and perhaps inherently, struggle to teach or to explain how they arrive at their solutions or why those solutions are superior. It is up to human beings to decipher the significance of what AI systems are doing and to develop interpretations. In some ways, AI is comparable to the

classical oracle of Delphi, which left to human beings the interpretation of its cryptic messages about human destiny.

If AI improves constantly—and there is no reason to think it will not—the changes it will impose on human life will be transformative. Here are but two illustrations: a macro-example from the field of global and national security, and a micro-example dealing with the potential role of AI in human relationships.

AI, Grand Strategy, and Security

In the nuclear age, strategy evolved around the concept of deterrence. Deterrence is predicated on the rationality of parties, and the premise that stability can be ensured by nuclear and other military deployments that can be neutralized only by deliberate acts leading to self-destruction; the likelihood of retaliation deters attack. Arms-control agreements with monitoring systems were developed in large part to avoid challenges from rogue states or false signals that might trigger a catastrophic response.

Hardly any of these strategic verities can be applied to a world in which AI plays a significant role in national security. If AI develops new weapons, strategies, and tactics by simulation and other clandestine methods, control becomes elusive, if not impossible. The premises of arms control based on disclosure will alter: Adversaries' ignorance of AI-developed configurations will become a strategic advantage—an advantage that would be sacrificed at a negotiating table where transparency as to capabilities is a prerequisite. The opacity (and also the speed) of the cyberworld may overwhelm current planning models.

The evolution of the arms-control regime taught us that grand strategy requires an understanding of the capabilities and military deployments of potential adversaries. But if more and more intelligence becomes opaque, how will policy makers understand the views and abilities of their adversaries and perhaps even allies? Will many different internets emerge or, in the end, only one? What will be the implications for cooperation? For confrontation? As AI becomes ubiquitous, new concepts for its security need to emerge. One of them is the capability to disconnect from the network on which it operates.

More pointed—and potentially more worrisome—issues loom. Does the existence of weapons of unknowable potency increase or decrease the likelihood of future conflict? In the face of the unknown, will fear increase the tendency to preempt? The incentives will be for opacity, which could mean absolute insecurity. In these circumstances, how will norms and rules for guiding and restraining strategy be established? The need to develop strategic concepts relevant to this new and inevitable technology has become overwhelming.

Human Contact

Google Home and Amazon's Alexa are digital assistants already installed in millions of homes and designed for daily conversation: They answer queries and offer advice that, especially to children, may seem intelligent, even wise. And they can become a solution to the abiding loneliness of the elderly, many of whom interact with these devices as friends.

The more data AI gathers and analyzes, the more precise it becomes, so devices such as these will learn their owners' preferences and take them into account in shaping their answers. And as they get "smarter," they will become more intimate companions. As a result, AI could induce humans to feel toward it emotions it is incapable of reciprocating.

Already, people rank their smartphones as their most important possession. They name their Roombas, and attribute intent to them where none exists. What happens when these devices become even more sophisticated? Will people become as attached to their digital pets as to their dogs—or perhaps even more so?

Societies will adopt these devices in ways most compatible with their cultures, in some cases accentuating cultural differences. In Japan, for example, as a result of both an aging population and Shintoism (which considers inanimate objects to have spirits not unlike humans'), AI companions may become even more widespread than in the West.

Given these developments, it is possible that in many parts of the world, from early childhood onward the primary sources of interaction and knowledge will be not parents, family members, friends, or teachers, but rather digital companions, whose constantly available interaction will yield both a learning bonanza and a privacy challenge. AI algorithms will help open new frontiers of knowledge, while at the same time narrowing information choices and enhancing the capacity to suppress new or challenging ideas. AI is able to remove obstacles of language and many inhibitions of culture. But the same technology also creates an unprecedented ability to constrain or shape the diffusion of information. The technological capacity of governments to monitor the behavior and movements of tens or hundreds of millions is likewise unprecedented. Even in the West, this quest can, in the name of harmony, become a slippery slope. Balancing the risks of aberrant behavior against limits on personal freedom—or even defining *aberrant*—will be a crucial challenge of the AI era.

The Future

Many public projections of AI have the attributes of science fiction. But in the real world, there are many hopeful trends. AI will make fundamental positive contributions in vital areas such as health, safety, and longevity.

Still, there remain areas of worrisome impact: in diminished inquisitiveness as humans entrust AI with an increasing share of the quest for knowledge; in diminished trust via inauthentic news and videos; in the new possibilities it opens for terrorism; in weakened democratic systems due to AI manipulation; and perhaps in a reduction of opportunities for human work due to automation.

As AI becomes ubiquitous, how will it be regulated? Monitored? As we enter a world where people are taught by AI, will there be the AI equivalent of "approved" school textbooks?

The challenge of absorbing this new technology into the values and practices of the existing culture has no precedent. The most comparable event was the transition from the medieval to the modern period. In the medieval period, people interpreted the universe as a creation of the divine and all its manifestations as emanations of divine will. When the unity of the Christian Church was broken, the question of what unifying concept could replace it arose. The answer finally emerged in what we now call the Age of Enlightenment; great philosophers replaced divine inspiration with reason,

experimentation, and a pragmatic approach. Other interpretations followed: philosophy of history; sociological interpretations of reality. But the phenomenon of a machine that assists—or possibly surpasses—humans in mental labor and helps to both predict and shape outcomes is unique in human history. The Enlightenment philosopher Immanuel Kant ascribed truth to the impact of the structure of the human mind on observed reality. AI's truth is more contingent and ambiguous; it modifies itself as it acquires and analyzes data.

How should we respond to the inevitable evolution it will impose on our understanding of truth and reality? The three of us have discussed many ideas: programming digital assistants to refuse to answer philosophical questions, especially about the bounds of reality; requiring human involvement in high-stakes pattern recognition, such as the reading of X-rays; developing simulations in which AI can practice defining for itself ambiguous human values—*what is ethical? reasonable? does no harm?*—in various situations; "auditing" AI and correcting it when it inaccurately emulates our values; establishing a new field, an "AI ethics," to facilitate thinking about the responsible administration of AI, the way bioethics has facilitated thinking about the responsible administration of biology and medicine. Importantly, all such efforts must be undertaken according to three time horizons: what we already know, what we are sure to discover in the near future, and what we are likely to discover when AI becomes widespread.

The three of us differ in the extent to which we are optimists about AI. But we agree that it is changing human knowledge, perception, and reality—and, in so doing, changing the course of human history. We seek to understand it and its consequences, and encourage others across disciplines to do the same.

We want to hear what you think about this article. Submit a letter to the editor or write to letters@theatlantic.com.

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

Eric Schmidt is the former CEO and chairman of Alphabet.

Daniel Huttenlocher is the founder and former dean and vice provost of Cornell Tech and the current dean of the MIT Schwarzman College of Computing.