CHATHAM HOUSE

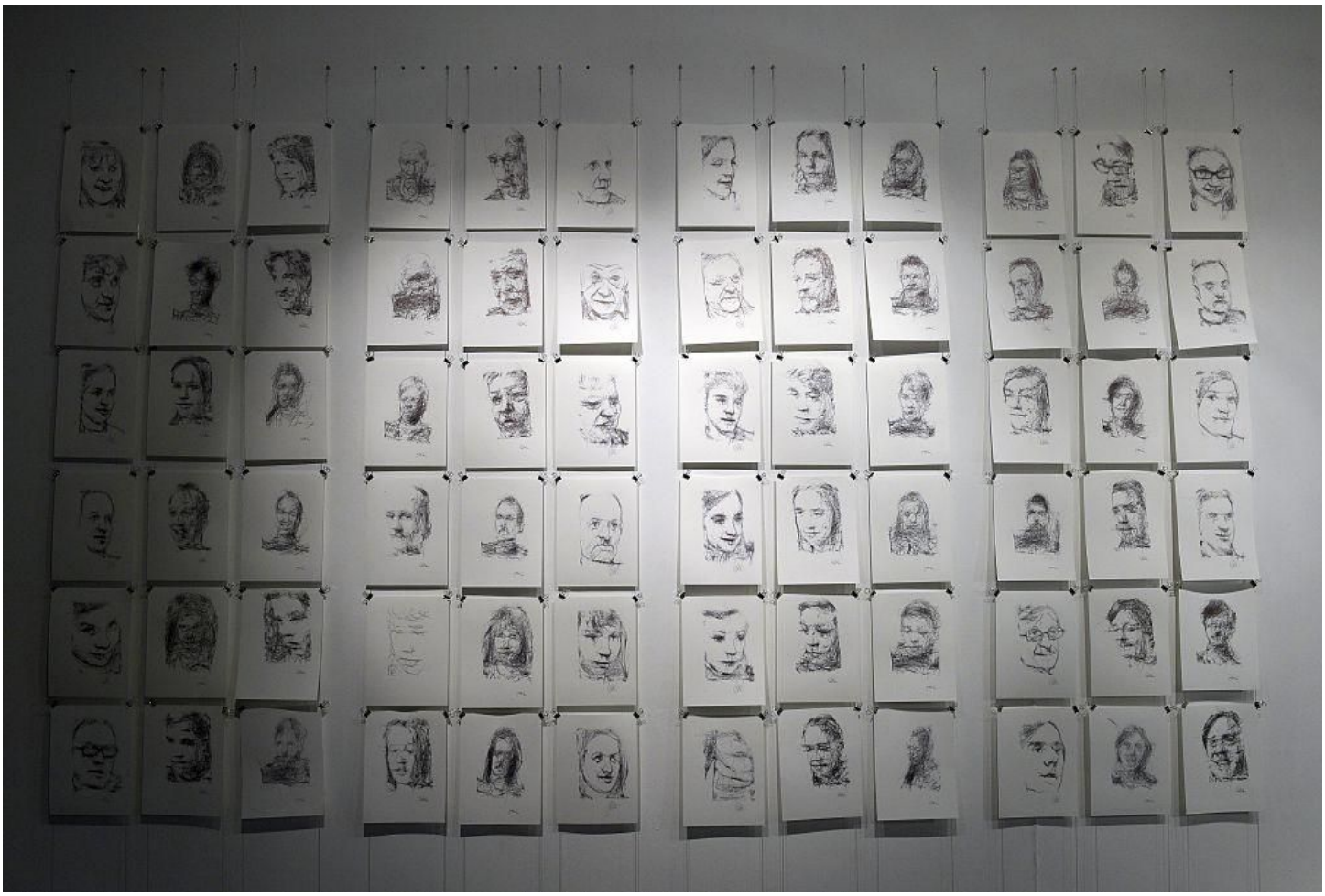
Combining AI's Power with Self-centered Human Nature Could Be Dangerous

Anthropomorphic bias assumes that as artificial intelligence develops, it will become more human-like. But the reality is likely to be far more disconcerting and incomprehensible.

EXPERT COMMENT 13 MARCH 2019 — 3 MINUTE READ

Jennifer Zhu Scott

Associate Fellow, Digital Society Initiative



—Portraits by '3 Robots Named Paul' - an art installation by Patrick Tresset which uses robots to sketch human models. Photo: Getty Images.

If we could shrink the entire history of our planet to one year, humans would have shown up roughly at 11pm on 31 Dec. In the grand scheme of things, we are insignificant. However, if we expand our thinking to the entire observable universe, our evolutionary success is a stroke of near-impossible luck that comprises all the biological conditions and chances required for us to become the dominant species on this planet. Of the 300 billion solar systems in the Milky Way, Earth is the only planet on which we know life exists. Out of the 8.7 billion known species on earth, we became the first general intelligence. Homo sapiens' evolutionary success is a miracle.

As far back as the human gene can record, Homo sapiens have been absorbing, accumulating, processing, learning, recording, and resolving infinite amount and types of data. Without pre-defined algorithms, we see, hear, taste, smell, touch, and think about everything that comes to us, all the time. The striking results of such learning form our physical features, cultures, languages, behaviours, histories, the world around us, and the ever-evolving capacity of our brain. In the 4 billion years of the Earth's history, the continuum of the last 200,000 years has created the most complex computer in our known universe. It sits above our shoulders, with up to 85 billion neurons constantly recording, calculating, processing from the moment we are born to the moment we die.

As the first general intelligence on this planet, our capabilities have evolved so much that today we found ourselves debating if we could artificially create synthetic general intelligence that would threaten our existence. Artificial general intelligence (AGI) has sparked perhaps the most fascinating and contentious debate among AI researchers, computer scientists, philosophers, neuroscientists, historians and anthropologists. There is no universal consensus on a detailed definition of AGI. Wikipedia states AGI 'is the intelligence of a machine that could successfully perform any intellectual task that a human being can'. However, the very definition of AGI is still debatable.

It is a necessary debate because the future of our species relies on how thoughtful we are today.

The first problem with human-like AGI is that we are still new in understanding our own brains. Our brain's physical implementations remain a mystery today. To build computing architectures that can perform all tasks that humans are capable of is physically impossible and economically unnecessary.

Besides, human capabilities never stop growing and evolving. At most, we could build many verticals of narrow intelligence that mimic human performance. If we can't replicate every possible task that human can perform, AGI, therefore, might never be human-like, self-aware, and conscious robots indistinguishable from humans, as we see in movies like *Ex Machina* or *The Terminator*. It is important to remember that movies or science fictions are art. The primary job of art is to provoke emotions, not to inform.

Secondly, most of the useful AI today are still 'single domain optimizers' — as Kaifu Lee accurately calls them. They are great at optimizing certain specific tasks but often useless with everything else. Such optimizers provide immediate economic value and therefore are most likely to be developed and adopted by the world. In a short to medium term, such 'stupid' AIs will prevail commercially.

So why do indistinguishable human-like robots come to our mind first?

The root of this phenomenon is the same reason fictional humanistic robotic characters in sci-fi movies can provoke our deep emotions. It is due to what scientists call 'anthropomorphic bias', meaning we project human characteristics to nonhuman creatures or objects. We have been doing this to animals and objects for as long as literature and legends go back. AI and robots are merely more recent targets.

Anthropomorphic bias, to me, is a peculiar form of narcissism. When it comes to intelligence and capabilities, many animals can outperform humans in specific tasks. Certain aspects of machine intelligence have also surpassed human intelligence decades ago. However, since we intuitively see the world through our lens, we unconsciously prefer nonhuman biological or synthetic intelligence to be like us. That is a distraction.

To properly understand and prepare what we are facing, we must go beyond such bias and challenge our perception on AGI.

DeepMind, arguably the most advanced AI company in the world, recently announced their AI AlphaStar had beaten the best human *StarCraft II* players, Dario 'TLO' Wünsch and Grzegorz 'MaNa' Komincz. AlphaStar is an elevation from AlphaGo0 that became the very best Go player within 40 days using Reinforcement Learning. AlphaGo, an earlier version of AlphaGo0, learned to play Go from all the recorded games with human champions and millions of simulations against itself. To initiate AlphaGo0's reinforcement self-learning, it was only given the basic rules of Go game and positive or negative feedback as it taught itself how to advance the craft.

AlphaStar is an impressive improvement because *StarCraft II*, a classic video game, requires multiple fast strategies formed with incomplete information, whereas with a Go game, players always have the complete information of every grid. Similar to AlphaGo0, AlphaStar has a very unorthodox style that baffled the world's best players.

The best minds of our species were surprised and perplexed by the ways how AlphaStar and AlphaGo 'think'. AlphaStar and AlphaGo are indeed 'artificial', but are far from being human-like. In 2017, Google's AI was also reported to have created languages that humans couldn't understand. Such 'inhuman' nature is a feature of the world's most advanced AIs. Assessing likely future scenarios, perhaps AGI won't be that 'general', instead, a combination or linkage of multiple 'super' intelligences far beyond human comprehension.

Such agents of synthetic superintelligence could be so strong that no human could understand its methods, or more disconcertingly, its intentions. It would be effortless and fast for such a superintelligence to expand its power to improve performance, but for humans to improve our biological brain power is a slow and challenging process. While self-improvement is a known feature, AI's self-perpetuity remains unproven. Humans, however, value self-existence and self-perpetuity. This is why stories about the afterlife and magic immortality pills exist in almost every culture.

Therefore, human-like robots taking over our world is a shallow and misleading distraction. The real threat for which we must prepare is the fatal combination between rapid self-improvement by synthetic superintelligence and self-centered human nature. After all, it is easier to create AGI by selecting and mixing what each side is best at. It may be a symbiosis, an augmented brain, an upload, a cyborg, or a combination. Who can set the boundary? What can we do today to prepare for this possibility? Those who pursue AGI must

bear the responsibility of clearly outlining the objectives of their endeavor, and fully assess the impact on humanity.

With AGI combined with self-centered human nature, the known knowns are that the incentive is enormous and the technical capacities will eventually catch up. The known unknown is the who, the when, and what the fatal combination will do with extraordinary power. The unknown unknowns? They are infinite, and that terrifies me.

Indeed, if our planet's history was one year, our species only showed up at the last hour. If we can live through midnight and welcome a new year, the moment will still be entirely in our hands.

Soon it won't be.

This article was originally published by Caixin Global.

Topics

DATA GOVERNANCE AND SECURITY FUTURE OF WORK TECHNOLOGY GOVERNANCE

Departments

ASIA-PACIFIC PROGRAMME

Related content

EVENT RECORDING

Watch: Technology and inclusion in the Asia-Pacific

EXPERT COMMENT

Collective action can spark innovation for data flows

28 JUNE 2021 — 3 MINUTE READ

RESEARCH PAPER

Artificial Intelligence for Healthcare: Insights from India

30 JULY 2020