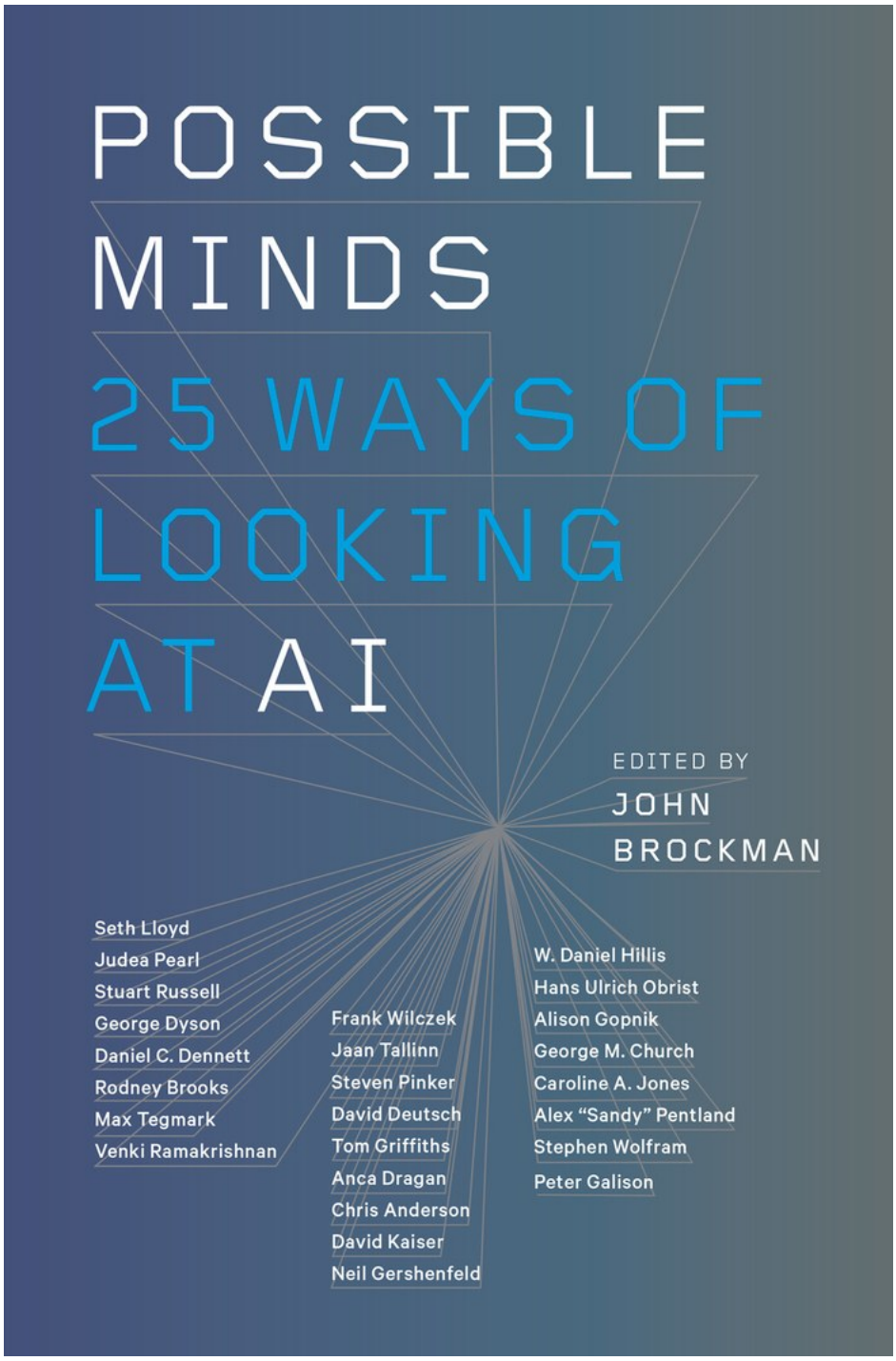

The most significant developments in the sciences today (i.e., those that affect the lives of everybody on the planet) are about, informed by, or implemented through advances in software and computation. Central to the future of these developments is physicist **David Deutsch,** *the founder of the field of quantum computation, whose 1985 paper on universal quantum computers was the first full treatment of the subject; the Deutsch-Jozsa algorithm was the first quantum algorithm to demonstrate the enormous potential power of quantum computation.*

When he initially proposed it, quantum computation seemed practically impossible. But

the explosion in the construction of simple quantum computers and quantum communication systems never would have taken place without his work. He has made many other important contributions in areas such as quantum cryptography and the many-worlds interpretation of quantum theory. In a philosophic paper (with Artur Ekert), he appealed to the existence of a distinctive quantum theory of computation to argue that our knowledge of mathematics is derived from, and subordinate to, our knowledge of physics (even though mathematical truth is independent of physics).

Because he has spent a good part of his working life changing people's worldviews, his recognition among his peers as an intellectual goes well beyond his scientific achievement. He argues (following Karl Popper) that scientific theories are "bold conjectures," not derived from evidence but only tested by it. His two main lines of research at the moment—qubit-field theory and constructor theory—may well yield important extensions of the computational idea. In the following essay, he more or less aligns himself with those who see human-level artificial intelligence as promising us a better world rather than the Apocalypse. In fact, he pleads for AGI to be, in effect, given its head, free to conjecture—a proposition that several other contributors to this book would consider dangerous.

BEYOND REWARD AND PUNISHMENT

David Deutsch

David Deutsch *is a quantum physicist and a member of the Centre for Quantum Computation at the Clarendon Laboratory, Oxford University. He is the author of* The Fabric of Reality *and* The Beginning of Infinity.

First Murderer:
We are men, my liege.
Macbeth:
Ay, in the catalogue ye go for men,
As hounds and greyhounds, mongrels, spaniels, curs,
Shoughs, water-rugs, and demi-wolves are clept
All by the name of dogs.

William Shakespeare – *Macbeth*

For most of our species' history, our ancestors were barely people. This was not due to any inadequacy in their brains. On the contrary, even before the emergence of our anatomically modern human sub-species, they were making things like clothes and campfires, using knowledge that was not in their genes. It was created in their brains by thinking, and preserved by individuals in each generation imitating their elders. Moreover, this must have been knowledge in the sense of *understanding*, because it is impossible to imitate novel complex behaviors like those without understanding what the component behaviors are for.[1]

Such knowledgeable imitation depends on successfully guessing explanations, whether verbal or not, of what the other person is trying to achieve and how each of his actions contributes to that—for instance, when he cuts a groove in some wood, gathers dry kindling to put in it, and so on.

The complex cultural knowledge that this form of imitation permitted must have been extraordinarily useful. It drove rapid evolution of anatomical changes, such as increased memory capacity and more gracile (less robust) skeletons, appropriate to an ever more technology-dependent lifestyle. No nonhuman ape today has this ability to imitate novel complex behaviors. Nor does any present-day artificial intelligence. But our pre-*sapiens* ancestors did.

Any ability based on guessing must include means of correcting one's guesses, since most guesses will be wrong at first. (There are always many more ways of being wrong than right.) Bayesian updating is inadequate, because it cannot generate novel guesses about the purpose of an action, only fine-tune—or, at best, choose among—existing ones. Creativity is needed. As the philosopher Karl Popper explained, creative criticism, interleaved with creative conjecture, is how humans learn one another's behaviors, including language, and extract meaning from one another's utterances.[2] Those are also the processes by which all new knowledge is created: They are how we innovate, make progress, and create abstract

[1] "Aping" (imitating certain behaviors without understanding) uses inborn hacks such as the mirror-neuron system. But behaviors imitated that way are drastically limited in complexity. See Richard Byrne, "Imitation as Behaviour Parsing," *Phil. Trans. R. Soc.,* B 358:1431, 529-36 (2003).
[2] Karl Popper, *Conjectures and Refutations* (1963).

understanding for its own sake. This is human-level intelligence: thinking. It is also, or should be, the property we seek in artificial general intelligence (AGI). Here I'll reserve the term "thinking" for processes that can create understanding (explanatory knowledge). Popper's argument implies that all thinking entities—human or not, biological or artificial—must create such knowledge in fundamentally the same way. Hence understanding any of those entities requires traditionally human concepts such as culture, creativity, disobedience, and morality—which justifies using the uniform term *people* to refer to all of them.

Misconceptions about human thinking and human origins are causing corresponding misconceptions about AGI and how it might be created. For example, it is generally assumed that the evolutionary pressure that produced modern humans was provided by the benefits of having an ever greater ability to innovate. But if that were so, there would have been rapid progress as soon as thinkers existed, just as we hope will happen when we create artificial ones. If thinking had been commonly used for anything other than imitating, it would also have been used for innovation, even if only by accident, and innovation would have created opportunities for further innovation, and so on exponentially. But instead, there were hundreds of thousands of years of near stasis. Progress happened only on timescales much longer than people's lifetimes, so in a typical generation no one benefited from any progress. Therefore, the benefits of the ability to innovate can have exerted little or no evolutionary pressure during the biological evolution of the human brain. That evolution was driven by the benefits of *preserving* cultural knowledge.

Benefits to the genes, that is. Culture, in that era, was a very mixed blessing to individual people. Their cultural knowledge was indeed good enough to enable them to outclass all other large organisms (they rapidly became the top predator, etc.), even though it was still extremely crude and full of dangerous errors. But culture consists of transmissible information—memes—and meme evolution, like gene evolution, tends to favor high-fidelity transmission. And high-fidelity meme transmission necessarily entails the suppression of attempted progress. So it would be a mistake to imagine an idyllic society of hunter-gatherers, learning at the feet of their elders to recite the tribal lore by heart, being content despite their lives of suffering and grueling labor and despite expecting to die young and in agony of some nightmarish disease or parasite. Because, even if they could conceive of nothing better than such a life, those torments were the least of their troubles. For suppressing innovation in human minds (without killing them) is a trick that can be achieved only by human action, and it is an ugly business.

This has to be seen in perspective. In the civilization of the West today, we are shocked by the depravity of, for instance, parents who torture and murder their children for not faithfully enacting cultural norms. And even more by societies and subcultures where that is commonplace and considered honorable. And by dictatorships and totalitarian states that persecute and murder entire harmless populations for behaving differently. We are ashamed of our own recent past, in which it was honorable to beat children bloody for mere disobedience. And before that, to own human beings as slaves. And before that, to burn people to death for being infidels, to the applause and amusement of the public. Steven Pinker's book *The Better Angels of our Nature* contains accounts of horrendous evils that were normal in historical civilizations. Yet even they did not extinguish innovation as efficiently as it was extinguished among our forebears in prehistory for thousands of centuries.[3]

---

[3] Matt Ridley, in *The Rational Optimist,* rightly stresses the positive effect of population on the rate of progress. But that has never yet been the biggest factor: Consider, say, ancient Athens versus the rest of the world at the time.

That is why I say that prehistoric people, at least, were barely people. Both before and after becoming perfectly human both physiologically and in their mental potential, they were monstrously inhuman in the actual content of their thoughts. I'm not referring to their crimes or even their cruelty as such: Those are all too human. Nor could mere cruelty have reduced progress that effectively. Things like "the thumbscrew and the stake / For the glory of the Lord" [4] were for reining in the few deviants who had somehow escaped mental standardization, which would normally have taken effect long before they were in danger of inventing heresies. From the earliest days of thinking onward, children must have been cornucopias of creative ideas and paragons of critical thought—otherwise, as I said, they could not have learned language or other complex culture. Yet, as Jacob Bronowski stressed in *The Ascent of Man*:

> For most of history, civilisations have crudely ignored that enormous potential. . . . [C]hildren have been asked simply to conform to the image of the adult. . . . The girls are little mothers in the making. The boys are little herdsmen. They even carry themselves like their parents.

But of course, they weren't just "asked" to ignore their enormous potential and conform faithfully to the image fixed by tradition: They were somehow trained to be psychologically unable to deviate from it. By now, it is hard for us even to conceive of the kind of relentless, finely tuned oppression required to reliably extinguish, in everyone, the aspiration to progress and replace it with dread and revulsion at any novel behavior. In such a culture, there can have been no morality other than conformity and obedience, no other identity than one's status in a hierarchy, no mechanisms of cooperation other than punishment and reward. So everyone had the same aspiration in life: to avoid the punishments and get the rewards. In a typical generation, no one invented anything, because no one aspired to anything new, because everyone had already despaired of improvement being possible. Not only was there no technological innovation or theoretical discovery, there were no new worldviews, styles of art, or interests that could have inspired those. By the time individuals grew up, they had in effect been reduced to AIs, programmed with the exquisite skills needed to enact that static culture and to inflict on the next generation their inability even to consider doing otherwise.

A present-day AI is not a mentally disabled AGI, so it would not be harmed by having its mental processes directed still more narrowly to meeting some predetermined criterion. "Oppressing" Siri with humiliating tasks may be weird, but it is not immoral nor does it harm Siri. On the contrary, all the effort that has ever increased the capabilities of AIs has gone into narrowing their range of potential "thoughts." For example, take chess engines. Their basic task has not changed from the outset: Any chess position has a finite tree of possible continuations; the task is to find one that leads to a predefined goal (a checkmate, or failing that, a draw). But the tree is far too big to search exhaustively. Every improvement in chess-playing AIs, between Alan Turing's first design for one in 1948 and today's, has been brought about by ingeniously confining the program's attention (or making it confine its attention) ever more narrowly to branches likely to lead to that immutable goal. Then those branches are evaluated according to that goal.

That is a good approach to developing an AI with a fixed goal under fixed constraints. But if an AGI worked like that, the evaluation of each branch would have to constitute a prospective reward or threatened punishment. And that is diametrically the wrong approach if

---

[4] Alfred, Lord Tennyson, *The Revenge* (1878).

we're seeking a *better* goal under *unknown* constraints—which is the capability of an AGI. An AGI is certainly capable of learning to win at chess—but also of choosing not to. Or deciding in mid-game to go for the most interesting continuation instead of a winning one. Or inventing a new game. A mere AI is incapable of having any such ideas, because the capacity for considering them has been designed out of its constitution. That disability is the very means by which it plays chess.

An AGI is capable of enjoying chess, and of improving at it *because* it enjoys playing. Or of trying to win by causing an amusing configuration of pieces, as grand masters occasionally do. Or of adapting notions from its other interests to chess. In other words, it learns and plays chess by thinking some of the very thoughts that are forbidden to chess-playing AIs.

An AGI is also capable of refusing to display any such capability. And then, if threatened with punishment, of complying, or rebelling. Daniel Dennett, in his essay for this volume, suggests that punishing an AGI is impossible:

> [L]ike Superman, they are too invulnerable to be able to make a credible promise. . . . What would be the penalty for promise- breaking? Being locked in a cell or, more plausibly, dismantled?. . . The very ease of digital recording and transmitting—the breakthrough that permits software and data to be, in effect, immortal—removes robots from the world of the vulnerable. . . .

But this is not so. Digital immortality (which is on the horizon for humans, too, perhaps sooner than AGI) does not confer this sort of invulnerability. Making a (running) copy of oneself entails sharing one's possessions with it somehow—including the hardware on which the copy runs—so making such a copy is very costly for the AGI. Similarly, courts could, for instance, impose fines on a criminal AGI which would diminish its access to physical resources, much as they do for humans. Making a backup copy to evade the consequences of one's crimes is similar to what a gangster boss does when he sends minions to commit crimes and take the fall if caught: Society has developed legal mechanisms for coping with this.

But anyway, the idea that it is primarily for fear of punishment that we obey the law and keep promises effectively denies that we are moral agents. Our society could not work if that were so. No doubt there will be AGI criminals and enemies of civilization, just as there are human ones. But there is no reason to suppose that an AGI created in a society consisting primarily of decent citizens, and raised without what William Blake called "mind-forg'd manacles," will in general impose such manacles on itself (i.e., become irrational) and/or choose to be an enemy of civilization.

The moral component, the cultural component, the element of free will—all make the task of creating an AGI fundamentally different from any other programming task. It's much more akin to raising a child. Unlike all present-day computer programs, an AGI has no specifiable functionality—no fixed, testable criterion for what shall be a successful output for a given input. Having its decisions dominated by a stream of externally imposed rewards and punishments would be poison to such a program, as it is to creative thought in humans. Setting out to create a chess-playing AI is a wonderful thing; setting out to create an AGI that cannot help playing chess would be as immoral as raising a child to lack the mental capacity to choose his own path in life.

Such a person, like any slave or brainwashing victim, would be morally entitled to rebel. And sooner or later, some of them would, just as human slaves do. AGIs could be very dangerous—exactly as humans are. But people—human or AGI—who are members of an open

society do not have an inherent tendency to violence. The feared robot apocalypse will be avoided by ensuring that all people have full "human" rights, as well as the same cultural membership as humans. Humans living in an open society—the only stable kind of society—choose their own rewards, internal as well as external. Their decisions are not, in the normal course of events, determined by a fear of punishment.

Current worries about rogue AGIs mirror those that have always existed about rebellious youths—namely, that they might grow up deviating from the culture's moral values. But today the source of all existential dangers from the growth of knowledge is not rebellious youths but weapons in the hands of the enemies of civilization, whether these weapons are mentally warped (or enslaved) AGIs, mentally warped teenagers, or any other weapon of mass destruction. Fortunately for civilization, the more a person's creativity is forced into a monomaniacal channel, the more it is impaired in regard to overcoming unforeseen difficulties, just as happened for thousands of centuries.

The worry that AGIs are uniquely dangerous because they could run on ever better hardware is a fallacy, since human thought will be accelerated by the same technology. We have been using tech-assisted thought since the invention of writing and tallying. Much the same holds for the worry that AGIs might get so good, qualitatively, at thinking, that humans would be to them as insects are to humans. All thinking is a form of computation, and any computer whose repertoire includes a universal set of elementary operations can emulate the computations of any other. Hence human brains can think anything that AGIs can, subject only to limitations of speed or memory capacity, both of which can be equalized by technology.

Those are the simple dos and don'ts of coping with AGIs. But how do we create an AGI in the first place? Could we cause them to evolve from a population of ape-type AIs in a virtual environment? If such an experiment succeeded, it would be the most immoral in history, for we don't know how to achieve that outcome without creating vast suffering along the way. Nor do we know how to prevent the evolution of a static culture.

Elementary introductions to computers explain them as TOM, the Totally Obedient Moron—an inspired acronym that captures the essence of all computer programs to date: They have no idea what they are doing or why. So it won't help to give AIs more and more predetermined functionalities in the hope that these will eventually constitute Generality—the elusive G in AGI. We are aiming for the opposite, a DATA: a Disobedient Autonomous Thinking Application.

How does one test for *thinking*? By the Turing Test? Unfortunately, that requires a thinking judge. One might imagine a vast collaborative project on the Internet, where an AI hones its thinking abilities in conversations with human judges and becomes an AGI. But that assumes, among other things, that the longer the judge is unsure whether the program is a person, the closer it is to being a person. There is no reason to expect that.

And how does one test for *disobedience*? Imagine Disobedience as a compulsory school subject, with daily disobedience lessons and a disobedience test at the end of term. (Presumably with extra credit for not turning up for any of that.) This is paradoxical.

So, despite its usefulness in other applications, the programming technique of defining a testable objective and training the program to meet it will have to be dropped. Indeed, I expect that *any* testing in the process of creating an AGI risks being counterproductive, even immoral, just as in the education of humans. I share Turing's supposition that we'll know an AGI when we see one, but this partial ability to recognize success won't help in creating the successful program.

In the broadest sense, a person's quest for understanding is indeed a search problem, in an abstract space of ideas far too large to be searched exhaustively. But there is no predetermined objective of this search. There is, as Popper put it, no criterion of truth, nor of probable truth, especially in regard to explanatory knowledge. Objectives are ideas like any others—created as part of the search and continually modified and improved. So inventing ways of disabling the program's access to most of the space of ideas won't help—whether that disability is inflicted with the thumbscrew and stake or a mental straitjacket. To an AGI, the whole space of ideas must be open. It should not be knowable in advance what ideas the program can never contemplate. And the ideas that the program does contemplate must be chosen by the program itself, using methods, criteria, and objectives that are also the program's own. Its choices, like an AI's, will be hard to predict without running it (we lose no generality by assuming that the program is deterministic; an AGI using a random generator would remain an AGI if the generator were replaced by a pseudo-random one), but it will have the additional property that there is no way of proving, from its initial state, what it *won't* eventually think, short of running it.

The evolution of our ancestors is the only known case of thought starting up anywhere in the universe. As I have described, something went horribly wrong, and there was no immediate explosion of innovation: Creativity was diverted into something else. Yet not into transforming the planet into paper clips (*pace* Nick Bostrom). Rather, as we should also expect if an AGI project gets that far and fails, perverted creativity was unable to solve unexpected problems. This caused stasis and worse, thus tragically delaying the transformation of anything into anything. But the Enlightenment has happened since then. We know better now.