AI'S NEW PROMISE

Diane Proudfoot and Jack Copeland
University of Canterbury, New Zealand

AI has always been full of big promises. Even back in the late 1940s and early 1950s, Alan Turing was predicting intelligent computers—machines that the average person is unable to distinguish from a human being on the basis of an email conversation concerning any topics whatsoever. When Turing suggested that, in order to build a "thinking machine", the machine be allowed to "roam the countryside" and learn by itself, his colleagues mocked the idea, saying "Turing is going to infest the countryside … with a robot which will live on twigs and scrap iron". Turing, on the other hand, was afraid that when his "child machines" were sent to school the human children would make "excessive fun" of them. Here he was no doubt speaking tongue-in-cheek, but he did say in 1952 that a machine *would* pass his now-famous test for intelligence—aka the "imitation game"—although only after "at least 100 years".

Many of AI's promises and claims in the 20th century were wildly over-optimistic. For instance, Herbert Simon and Alan Newell announced in 1958 that "there are now in the world machines that think, that learn, and that create"; and in the 1980s Roger Schank's company Cognitive Systems, Inc., claimed that their programs have "the same kind of knowledge that people use … [and] understand a sentence just the way a person does". It was Schank's confidence that provoked John Searle's famous "Chinese room argument" against the possibility of "strong AI"—the hypothesis that, solely by executing a program that a simple Turing machine could run, a computer really can do things like understand a sentence just the way a human being does.

In recent years AI researchers have been making the most significant promise yet—*immortality* for the human race. In 2001 Ray Kurzweil (author of *The Age of Spiritual Machines* and *The Singularity is Near*) predicted "the merger of biological and nonbiological intelligence" and "immortal software-based humans"—all within a few decades from now. First, humans will be "enhanced", using biotechnology and nanotechnology; "By the 2030s, we will be more non-biological than biological", Kurzweil claims. Later, according not only to Kurzweil but to other "technological futurists", such as Nick Bostrom, Ben Goertzel, Hans Moravec, and Frank Tipler (author of *The Physics of Immortality*), we will be entirely non-biological—"posthumans", "ex-humans", or "postbiologicals". According to these theorists, the program that comprises your mind can be uploaded into a computer, and when your biological body self-destructs from disease, accident, or old age, you can live on. Your posthuman life may be led "out on the Web" or in a new cybernetic body, as you choose.

Being software-based has many advantages, futurists say, including being able to "transmit oneself as information at the speed of light" and "think a thousand times faster". As a consequence of acquiring cognitive and affective capabilities that humans now can "only dream about", you will, it is claimed, experience "surpassing bliss" and have a "truly meaningful" life. According to Kurzweil, this "freeing of the human mind from its severe physical limitations of scope and duration" is "the necessary next step in evolution". Most importantly, humans will escape death—with regular backups, forever. Immortality won't be boring, as some philosophers have feared, since as an upload you can experience your "ideal fantasy worlds"—and if you were to become bored, you could in any case simply alter your own programming so that you returned to a blissful state! To achieve immortality, futurists say, all *you* need do is survive until radical life-extension technologies are developed. (In his book *Fantastic Voyage: Live Long Enough to Live Forever* Kurzweil recommends a programme of diet, exercise, aggressive nutritional supplementation, and stress-management; he also, with Terry Grossman, offers "Ray and Terry's Longevity Products".) But even if you perish before then, help is at hand. According to many futurists, the superhuman-level artificial intelligences just around the corner will be able to recreate your program after your "death", and resurrect you—along with every other human being who ever lived.

Are these claims science or fantasy? According to Turing, "conjectures" are important to scientific research, which, contrary to popular opinion, doesn't "proceed inexorably from well-established fact to well-established fact". To avoid harm, though, it must be made clear "which are proved facts and which conjectures", he said. Futurists write as though a posthuman future is highly probable, but their hypothesis that we will—or even could—become software-based postbiologicals is almost entirely conjecture. From a technological point of view, nobody has even the beginnings of an idea how the human mind might be "uploaded". Kurzweil and like-minded futurists claim that soon we will be able to use high-resolution scanning in order to reverse engineer the human brain, discover "the software of intelligence", and simulate a specific brain in a computer. But even if imaging technology greatly improves, *what* do we scan and try to simulate, in order to capture the mind? Neurons? Micro-tubules? Cells? Atoms? Elementary particles? Kurzweil says confidently, "We ultimately will be able to capture and recreate [a human being's] pattern of salient neural and physical details to any desired degree of accuracy"—but *which* details are "salient" and what degree of accuracy is "necessary"? No-one knows. "No matter how detailed a picture you had of the inside of a brain", AI researcher Drew McDermott points out, "you wouldn't know which details were important". Moreover, when futurists talk of what is "salient" or "necessary" in simulating the brain, usually they mean what is relevant to replicating the *computations* that they assume are performed by the brain. Futurists typically take for granted that brain activity is Turing-computational—i.e., can be replicated in a Turing machine—and it is these computations that are to comprise "the software of intelligence". But *is* the brain a computational device in this sense? Or is it perhaps a "hypercomputer", or even a spongy swamp of chemical interactions that largely defy description as computations of any type? Again no one knows.

Even some of those researchers in AI who are optimistic that we can develop a computational theory of intelligence and even that we will build thinking machines, challenge Kurzweil's extrapolations from past technological progress—and in general the futurists' methods and time-frame for achieving human-level and superhuman-level AI. McDermott emphasizes the danger of AI's "overhyping" and thereby shooting itself in the foot. There has been remarkable progress in AI since Turing talked of building a machine that would "emulate the brain", but it hasn't been (at least directly) in brain-building. In fact many researchers have abandoned the traditional goal of human-level AI. Advances have been in what is sometimes called "narrow AI"—task-specific devices such as medical diagnostic programs, voice-recognition software, and search engines like Google. It is a *very* long way from these machines to human-level AI (or "artificial general intelligence")—let alone to human "uploads".

Even if the theory and technology of simulating the human brain were perfected, the notorious philosophical question remains: would this be to simulate the *mind*? There are numerous things that a computer simply might not be able to do—fall in love, or enjoy strawberries and cream, to pick two examples that Turing discussed. It's possible to claim that a machine *thinks*—i.e. has a mind—even if it doesn't have the capacity for emotions or sensations. Even if we granted this, though, would you still be *you*, if your ability to enjoy dessert or fall in love were eliminated in the process of uploading? For *if* it is true that a computer cannot fall in love or enjoy strawberries and cream, then no more will your upload be able to. But futurists typically see no problem here; they assume that, once we simulate a specific human brain computationally, the simulation will inherit the human being's capacity for thoughts, emotions, and sensations. They acknowledge that humans are *embodied* creatures, and so allow that the simulation might need to interact with a specific body, if it is to be a human mind—but again there's no difficulty, they say, since an upload can be implemented in an artificial brain in an artificial body, or be provided with a virtual body in a virtual environment. Either way, futurists imply, the upload will be as capable as you of enjoying strawberries and cream. This whole way of theorizing, however, involves betting on a particular, computational, solution to the mind-body problem—a solution that may in time come to seem as deeply wrong as the ancient idea that the function of the brain is to cool the blood.

Even if the simulation of your brain were a fully-fledged thinking and feeling mind, we can still ask: would it be *your* mind? Is the upload really *you*, as futurists claim? According to Kurzweil, we can verify that the upload is you by means of a customized Turing test—by, he says, "convincing a human judge that the uploaded re-creation is indistinguishable from the original specific person". But even if an upload were to do well in a "Ray Kurzweil" imitation game, this would not show that the upload is Kurzweil. An actor might (under financial inducement) spend the last ten years of his life studying Kurzweil; his upload takes the Ray Kurzweil Turing test—and passes, but thanks only to the actor's skill at imitating Kurzweil. To assist with the test, the upload might even have accessed memories contained in the genuine Kurzweil-simulation (which has been made freely available on the Internet). Linking survival of self to a customized Turing test is a nice idea, but it isn't going to work.

To understand why some people object to the claim that a simulation of your brain is you, suppose that you have signed on to be uploaded once your natural body dies but something goes wrong with the IT company's server, and the you-simulation boots up while you're still very much alive. This upload *can't* be you, it seems—*you're* the one complaining to the company about their poor service, and demanding that they delete the upload. And if this upload isn't you *before* you die, then why should it be you *after* you die? Or suppose, on the other hand, that it's not the server's timing that goes haywire, but its backup system. The server correctly switches on the you-simulation after your death, but unfortunately it simultaneously switches on the backup of the you-simulation. Now there are *two* uploads—so are there two *yous*? These two "yous" could even be doing different things at the same time, for example upload-1 opting for a life of licentiousness on the Web while upload-2 joins a virtual holy order dedicated to prayer and spiritual advancement. In this case, is *one* person (you) leading literally *two* different lives at the same time? That looks like a contradiction. And if a year later upload-1 has acquired a robot body and lives in Slough, selling conventional life insurance to people too poor to pay to be uploaded, whereas upload-2 has shifted to a mainframe in a Pondicherry ashram and become a world-famous astrologer, are they *still* both you?

Many futurists have great faith in the superhuman abilities of future artificial intelligences, and it's possible that these AIs will carefully supervise the system to prevent there ever being two copies of you at one and the same time. Though technology companies of the future might resent this strict policing—they might argue, with some justice, that a single copy of the you-simulation could too easily be destroyed, for example in a fire or an earthquake. Moreover, their sales literature might claim, by running several copies of the you-simulation at the same time, you would find multi-tasking very easy—and you would have greatly increased opportunies for bliss. But all these scenarios highlight the philosophical problem at the heart of AI's new promise. How *can* there possibly be more than one you at the same time, even if in actual fact (thanks to careful management of the technology) there is only ever one?

This is what is known in the trade as the *duplication* problem, and the problem cases we have been discussing seem to suggest that a simulation of your brain is (at most) merely a *replica*—it isn't really *you*. But in fact the issue is wide open, and philosophers are divided over whether there is a way of defusing the duplication problem. In our view, one way of tackling it is to abandon the idea that statements like "That upload is you" are *determinately* true or false. Derek Parfit, writing about earlier forms of the duplication problem—involving ingenious brain surgeries and glitchy "teletransportation", rather than computer simulation—suggested a version of this approach. He imagined the case where his own brain is split in two, with one half being successfully transplanted into the body of one of his brothers, and the other half into another brother's body (the three brothers are identical triplets, and Parfit's body and his brothers' brains have been fatally injured). Parfit said, "We might claim that, to the question, 'Would I be either of the resulting people?', there is no true answer".

Adapting this strategy to the hypothetical case where the server generates two you-simulations at the one time, the question "Is either upload-1 or upload-2 *you*?" is simply a question with no true answer. This way, the problem cases seem no longer to lead to contradiction, because it isn't *true* that you are leading two different lives at the same time—but the cost is that, by the same token, nor is it true that something identical to your mind still exists. Our own preferred approach is (as we explain elsewhere) to investigate beefier versions of the idea that identity is indeterminate—versions where saying that the statement "That upload is you" is indeterminate is to say something a little more informative than simply that the statement is neither true nor false, and where indeterminacy does not automatically rule out its being *true* that something identical to your mind still exists.

What might Turing have said about the duplication problem? Impossible to know, but he would undoubtedly have relished the modern debate about AI's future. He said it was probable that, once artificial intelligence emerged, "it would not take long to outstrip our feeble powers". He continued, "There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control". Nor, Turing believed, would there be any "particularly human characteristic" that machines could not imitate. How would humans feel about this? Turing speculated, "Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled. A similar danger and humiliation threatens us from the possibility that we might be superseded by the pig or the rat". Some AI researchers today take an even bleaker view. There are those who regard a "posthuman" future with foreboding—fearful, as Bill Joy says, of the "power of destructive self-replication" and of a non-biological existence in which "our humanity may well be lost". We should not "pursue near immortality without considering the costs, without considering the commensurate increase in the risk of extinction", Joy solemnly cautions us. Surpassing bliss, or extinction? At this point in time, it's a case of picking your fantasy.