Just Who Will Be We, in 2493?

Douglas R. Hofstadter
Director, Center for Research on Concepts and Cognition
Indiana University, Bloomington
2003

I do research on computer modeling of the human mind's workings, sometimes called "artificial intelligence", or "AI". Though this is a discipline many consider futuristic, I am not a futurologist. By contrast, some of my colleagues have occasionally had the hubris of predicting a rosy short-term future for AI, and failed dismally. Herbert Simon, for instance, one of AI's founders, is notorious for predicting in 1958 that a computer would be world chess champion within a decade. As we know in 1996, Simon was quite wrong.

But with the recent computer-versus-human chess match, I was stimulated to ponder about where things are going. If the past is any guide to the future, the look-ahead horizon of brute-force chess programs will slowly increase, as it has been doing with regularity for the past 40 years or so, and at some point will encroach upon the further-out and blurrier horizon that so-called "strategic" or "positional" play affords human grandmasters. At that point, machine play will start to look creative, insightful, brilliant. So far, of course, we are not there. Indeed, I have heard Deep Blue's performance in its final two games against Kasparov variously described as "meandering", "aimless", "confused", and "below even Master level". Are we humans therefore to exult in our triumph over the alien silicon monster? And what about the year 2008, when Deep Blue's successor, Deep Pink, finally overtakes the future world champ Judith Polgar — should we then shed bitter tears?

A word of perspective. The way brute-force chess programs work doesn't bear the slightest resemblance to genuine human thinking, and so for me they have little intellectual interest, although I enjoy the spectacle of the battle. My research is focused on what I consider the core of human intelligence — the ability to adapt to different types of domain, the ability to spot the gist of situations amidst a welter of superficial distractors, the ability to see abstract resemblances between situations that on their surface seem completely disparate, the ability to be reminded by one situation of another one that is tremendously different at first glance and yet amazingly analogous at a deeper level. Spotting hidden patterns, extracting deep gists, forming high abstractions, making subtle analogies — these to me define the crux of the mental; they are what we do best of all creatures, natural and artificial, on the surface of this tiny huge green ball spinning its way through vast empty chasms of space.

Do machines do these kinds of things yet? No, not very well. Are they on their way to doing so? Well, to a degree. Their capacities at these tasks are still very rudimentary, but we are making progress. For decades, cognitive and perceptual psychologists have been devising beautiful and deep probes into the mechanisms behind human memory, the associative structures that underlie words and concepts, the way that perception hooks seamlessly into cognition, and so forth — and their findings are not going unheeded by the folks working on computer models of the mind. My own research on creative analogy-making by computer, for instance, is pervasively influenced by such studies.

So if we shift our attention from the flashy but inflexible kinds of game-playing programs like Deep Blue to the less glamorous but more humanlike programs that model analogy-making, learning, memory, and so forth, being developed by cognitive scientists around the world, we might ask, "Will *this* kind of program ever approach a human level of intelligence?" I frankly do not know. Certainly it's not just around the corner. But I see no reason why, in principle, humanlike thought, consciousness, and emotionality could not be the outcome of very complex processes taking place in a chemical substrate different from the one that happens, for historical reasons, to underlie our species.

The question then arises — a very hypothetical one, to be sure, but an interesting one to ponder: When these "creatures" (why not use that term?) come into existence, will they be threats to our own species? My answer is, it all depends. What it depends on, for me, comes down to one word: *benevolence.* If robot/computers someday roam along with us across the surface of our planet, and if they compose music and write poetry and come up with droll jokes — *and* if they leave us pretty much alone, or even help us achieve our goals — then why should we feel threatened? Obviously, if they start trying to push us out of our houses or to enslave us, that's another matter and we should feel threatened and should fight back.

But just suppose that we somehow managed to produce a friendly breed of silicon-based robots that shared much of our language and culture, although with differences, of course. There would naturally be a kind of rivalry between our different types, perhaps like that between different nations or races or sexes. But when the chips were down, when push came to shove, with whom would we feel allegiance? What, indeed, would the word "we" actually mean?

There is an old joke about the Lone Ranger and his sidekick Tonto one day finding themselves surrounded by a shrieking and whooping band of Indians circling in on them with tomahawks held high. The Lone Ranger turns to his faithful pal and says, "Looks like we're done for, Tonto..." To which Tonto replies, "What do you mean, *we,* white man?"

Let me suggest a curious scenario. Suppose we and our artificial progeny had coexisted for a while on our common globe, when one day some weird strain of microbes arose out of the blue, attacking carbon-based life with a viciousness that made today's Ebola virus and the old days' Black Plague seem like long-lost friends. After but a few months, the entire human race is utterly wiped out, yet our silicon cousins are untouched. After shedding metaphorical tears over our disappearance, they then go on doing their thing — composing haunting songs (influenced by Mozart, the Beatles, and Rachmaninoff), writing searching novels (in English and other human languages), making hilarious jokes (maybe even ethnic and sexual ones), and so on. If we today could look into some crystal ball and see that bizarre future, would we not thank our lucky stars that we had somehow managed, by hook or by crook, to propagate ourselves into the indefinite future by means of a switchover in chemical substrate? Would we not feel, looking into that crystal ball, that "we" were still somehow alive, still somehow *there*? Or — would those silicon-chip creatures bred of our own fancy still be unworthy of being labeled "we" by us?

For whom would we root, then, if in peering into the crystal ball, we witnessed a *carbon-based* race of alien invaders from some planet of Betelgeuse approaching the earth and systematically trying to wipe out the gentle, benevolent race of silicon-based intelligences to which we humans had intellectually but not biologically given rise? For that matter, forget the space aliens and simply imagine an all-out battle for survival

between our hypothetical benevolent silicon-based robots and some aggressive new carbon-based form of life — some mutant form of giant wasps or spiders, say, or even a mercilessly belligerent chimpanzee society — that arose on earth itself. Would we as humans watching helplessly on the sidelines be indifferent to the outcome? Or would we mindlessly align ourselves with the carbon? Or — is it within our ability to extend the word "we" from our simple *genetic* pool to something more abstract, something based on a certain way of thinking and feeling and caring, irrespective of the physical medium in which it is embedded?

I once gave a lecture in Holland in which I suggested such a vision of benevolent silicon creatures and suggested that the word "we" might someday come to encompass *them*, just as it now encompasses females and males, old and young, yellow and red, black and white, gay and straight, Arabs and Jews, weak and strong, cowardly and brave, short and tall, clever and silly, and so on. The next speaker, a gentle-looking, eloquent elderly fellow — indeed, quite resembling benevolent old Einstein — responded by arguing vociferously that the mere act of trying to develop artificial intelligence was inherently dangerous and evil, and that we should never, ever let computer programs make moral judgments, no matter how complex, subtle, or autonomous the programs might be. He argued that computers, robots, whatever they might become, irrespective of their natures, *must* in principle be kept out of certain areas of life — that our species has an exclusive and sacred right to certain behaviors and ideas, and this right must be protected above all.

Well, to my deep astonishment, when this gentleman had finished his pronouncements, nearly the entire audience rose to its feet and clapped wildly. Dazed, I could not help but be reminded of the crudest forms of racist, sexist, and nationalist oratory. Despite its high-toned and moralistic-seeming veneer, this exhortation and the audience's knee-jerk reaction seemed to me to be nothing more than a mindless and cruel biological tribalism rearing its ugly head. And this reaction, mind you, was in the supremely cosmopolitan, anti-Fascistic, internationally-minded country of Holland! Can you imagine how my ideas would have been greeted in the Bible Belt, or in Teheran or the Vatican?

Why can we humans not open our little minds to the potential idea that if, in some wild and far-off day, we finally succeeded in collectively creating nonbiological creatures that perceived, that enjoyed, that suffered, that joked, that loved, that hated, that even created, the very word "we" would at that very moment have opened up its semantic field to embrace these, the products of our hearts' deepest yearnings?

Why should a race of benevolent if fleshless beings be any less worthy of being considered our "children's children" than the potential gang of murderers and rapists that might spring forth from the union of my sperms with your ova? Just how far out does the circle stretch, whose radius is defined by the slippery word "we"? I wonder.