Future AI: Unhuman, Superhuman, Humane

In coming decades, computer engineers will build an entity with intelligence at a level able to compete with human intelligence. This entity will want to improve its intelligence, and will be able to do so. The process of improvement will repeat itself, until the artificial intelligence reaches levels far above the human; it will then achieve its goals easily. For the sake of the human future, the AI's goals must be defined, before the AI is built, to benefit humanity.

There are those who doubt that artificial intelligence at a roughly human level is possible within a few decades. Douglas Hofstadter, the author of *Gödel Escher Bach*, claims that centuries will pass until we reach this milestone. Hofstadter sees in humanity deep, complex, and even mysterious features which he thinks are impossible in an AI. But though the human mind is indeed more complex than any other object known in the universe, it is in the end nothing more than a protein-based machine, built of neurons. As a physical entity, there is nothing in it that cannot, in principle, be copied, even if this copying is difficult in practice.

But copying the brain's precise operation is not the essential thing. What we really care about is the effect of this AI on our world and on our future; the AI's ability to act in human-like ways is secondary. Alan Turing, the founder of AI, understood this from the first: In his seminal article, in which he described his famous imitation test, he stated that imitation is just a benchmark, a point of comparison which is convenient so long as we lack more precise measures.

A better understanding of intelligence will focus on the ability to achieve desired goals in general, and not just on the goal of mimicking humans. Thus, we can define intelligence as the ability to achieve complex goals in complex environments, with an emphasis on efficiency for goals and environments which are relevant to us humans. Mathematically, intelligence is optimization power in the broadest sense, in other words, the ability to maximize utility functions. With this approach to intelligence, we can look more clearly into the range of possibilities ahead of us: There may well be extremely non-anthropomorphic intelligences, those which do not necessarily resemble human minds. This is particularly true of AIs which are not close imitations of the functioning of the brain, but rather based on with algorithms and data structures designed and developed from scratch, an approach called Artificial General Intelligence (AGI).

Because of our evolutionary origins, all humans have a similar mental structure. Sexual reproduction with its mixing of genes does not allow much variation in a population. It may appear different people think very differently, but we must remember that none of us has ever seen a human-level intelligence which is not human. So, in trying to imagine a different intelligence, we must set aside our intuitions on the topic. By focusing on optimization, rather than on the intelligence we know, with all its baggage, we can separate the core of the concept from hidden assumptions based on our experience with people.

In this article I will focus primarily on future AGIs, those with a designed architecture, rather than on brain emulations (uploads). There are two reasons for this: First, the characteristics of brain emulations are already understood to us, because we are familiar with human behavior. In contrast, AGI based on theoretical foundations opens to us many more possibilities that are worth investigating. Second, AGI has advantages in the race to build a new general intelligence, as compared to complex andhard-to-understand biological structures. The first airplane was built on aerodynamic principles, with an indirect inspiration from birds, not as a precise emulation of birds with feathers and flapping wings.

When?

Is it even possible that engineers will build a non-human general intelligence in the next few decades, whether a brain emulation or based on theoretical foundations? According to some forecasts, and in particular those of Ray Kurzweil (*The Singularity is Near*, 2005), this milestone will be achieved by 2030.

There are some trends which support this possibility. The computational power in hardware that $1000 (inflation-adjusted) can purchase has been increasing exponentially over the last century. Algorithms are also improving. For many tasks in narrow AI (the type we have today), improvements in the algorithms are contributing to the increase in computational speed, even more than hardware improvements. At the same time, neuroscientists are learning how the brain works: Scanning technologies are also improving exponentially over the years, and today, scientists know a lot about the function of many parts of the brain, although the mechanism which builds thoughts out of electrochemical neural impulses is not yet understood.

All these facts suggest that a non-human general intelligence can be built within a few decades. But technological acceleration is neither necessary nor sufficient. In order to build such an entity, scientific breakthroughs will be needed. Although all inventors depend on earlier inventions, the breakthroughs will depend primarily on the genius and hard work of a handful of scientists and engineers, rather than on on technological and scientific trends. It is impossible to prophesy when these breakthroughs will occur, but whether they come in a few years or a few centuries, we must think deeply about this leap from human intelligence to a unhuman and potentially superhuman type of intelligence, because of its critical implications.

Goal-Seeking

The definition of intelligence as the ability to achieve goals may seem unsuited to characterizing the human mind. After all, people are driven by a mix of multiple conflicting desires, which are unclear to a person's own consciousness, and which frequently change. This situation is a product of our goals' origins as instruments for evolution's "goal," which is to make as many copies as possible of a gene. This is why humans strive to gather social status, to learn, to have sex, to survive, and much more. Of course, humans do not serve evolution consciously. Because of the many changes in our environment over the millennia, humans have long since deviated from evolution's goals, as illustrated by birth control and overeating. But incoherent though

this mix of desires is, it still constitutes a goal-system. Humans do not behave randomly, but rather work in certain directions, which are quite similar, though certainly not identical, in all humans.

An artificial intelligence might also work towards a complex and confused goal system, particularly if it is built on the human example. But if an AI is built by engineers who want it to solve a certain problem, its goal system may well be much simpler, for example, “Increase the amount of money in this hedge fund as much as possible,” “defeat our enemy in war,” or “find a cure for cancer.”

This intelligent entity, once created, will want to improve its own intelligence. This is because an improvement can contribute to achieving the entity’s end goals.

The entity will not only want to, but also *be able* to improve its own intelligence. An AI running on computer hardware can add processors and memory, or analyze and optimize its own source code. It could build copies of itself, or build entities of a new generation, which do not resemble it except in contributing to the achievement of the same goal. Having improved itself, it will be better able to improve itself more effectively, yet again.

These are techniques which are not available to humans. We can accumulate knowledge and learn new skills for self-improvement, but not change the structure of our brains. And when we humans create a new generation, our children do not work to achieve our precise preferences!

On the assumption that engineers will bring the AGI to roughly human levels, there is no reason that the improvement has to end precisely at that point. Humans have the least possible intelligence needed to create a civilization. There is nothing special about this level; it is not a necessary stopping point. Every iteration of improvement will add more ability until the intelligence reaches levels which are far above ours.

A superintelligence is very likely to achieve its goals. If we want it to work for the benefit of humanity and not to our detriment, it has to *want* our well-being. We will not be able to stop an intelligence which is smarter than us, nor to imprison it in a computer, nor to dictate laws which constrain its behavior. It will bypass with ease, if it so wants, any trick which we can think up to stop it.

It is very difficult to define goals whose fulfilment will benefit humanity. Our goal-system is variegated, and the omission of a small part of this mixture will bring disaster to humanity.

For instance, we can imagine a superintelligence which is working towards the exalted goal of “world peace.”. And it can do this relatively easily by instantly extinguishing all human life. A very bad outcome! If its goal is to maximize the love in the world, it can do so by changing our brain so that we feel nothing but love, like a mother rabbit for its kits, while also taking away initiative and ambition, which so often bring conflict and strife.

Goals such as “peace” and “love” arouse strong positive feelings, so it is problematic to focus our discussion on them. But even bland goals, with no emotional content,

endanger the future of the human species. The standard thought experiment involves paperclips. A powerful optimizer with the goal “collect as many paperclips as possible” will take various measures to maximize the number of paperclips in its possession. Self-improvement will help in its mission. When it reaches levels of power far above that of the entire human race, we humans will not have much influence on the achievements of its goal, one way or the other. Converting all the material in our planet to paperclips will help as it works towards maximizing its paperclip collection. More generally, using up all the resources which are necessary for people to live and thrive, including our planet’s atoms, will bring about the extinction of our species.

This behavior seems more like super-stupidity than super-intelligence. Yet we are not dealing with intelligences in the human sense, but rather more generally with optimizers, entities which try to maximize utility functions. The function “number of paperclips in my collection,” and in fact *most* possible utility functions, do not benefit humanity.

The danger to humanity is not malice or rebelliousness on the part the AI. The desire to throw off the yoke in humans comes from our evolutionary background. Though choosing one’s own way is usually valuable for achieving goals, this is not the case if benefitting humans is central to the AI’s aims.

Instead, the risk is from a secondary effect of successful optimization towards goals which we set the AI, if the goal system does not include *everything* that matters to us. The omission of any one of our many values, including love in all its varieties, freedom, health, physical activity, learning, aesthetics, creativity, or anything else, risks bringing about a world what we do not want to live in.

Thus, the AGI which we create must be *humane*: It must seek the full set of values of all humans. Not for itself, of course--I’m not talking about an AGI which is built to fall in love or to get a powerful aesthetic thrill from a well-baked torte--but rather in helping humans achieve what they truly want. If a superintelligence applies its abilities towards human well-being as its goal, then we will have a very good future ahead of us.

Intuition Breakers

These extreme effects of a powerful general optimizer are hard to grasp, given our intuitions, which are based on the human example. There are no such AGIs today which we can use as examples to familiarize ourselves with super-powerful goal-seeking. There are, however, a few powerful optimizers in existence which we can use to help us understand that not all optimizers need have goals or architecture or results resembling those of the human brain.

Evolution is an abstract concept representing optimization towards the maximal number of copies of a gene in a given environment by means of natural selection. Evolution is quite unhuman: It has no physical form, no consciousness, no modeling capabilities; it has no memory or predictive capabilities; but evolution is superhuman in some areas: it created oaks, fish, tigers, and humans. Evolution is not just superhuman and unhuman, but also inhumane: It created aging, diseases, parasites, and other forms of suffering which would seem like sheer evil if a human had created them.

Markets are another type of powerful optimizer. A market maximizes the joint benefit of all buyers and sellers active in it. They are far from perfect, as seen in recent market crashes, but they superhuman in that they produce a better outcome than one human or a small ground of humans who make economic decisions for everyone, like the Gosplan government bureau which attempted to control the Soviet economy. Markets, like evolution, are unhuman, with no physical form, no consciousness; they can sometimes be inhumane as well, with no conscience nor feelings whatsoever, beyond those of the individuals participating in the market.

Another type of superhuman and unhuman optimizer is purely theoretical: The perfect optimizer. Imagine a hypothetical entity which *instantly* achieves, *any* goal set for it, without any of the complex mix of drives, emotions, self-interest, and moral compunctions of human.

To formalize this concept, the mathematical model AIXI has been developed. It works in a framework in which an agent interacts with its environment over multiple rounds of observation, reward, and action, trying to get the greatest possible reward. The reward which the agent receives is a function of the environment and the agent's history of interaction with its environment, but the agent doesn't know the function in advance. It has to learn from observing the results of its actions.

AIXI achieves rewards provably greater than any other agent, across the full range of environments and reward functions, as follows: On each round of interaction, AIXI extrapolates the mathematical expectation of reward for *all* possible algorithms far into the future, and chooses the best one. Because there are infinitely many algorithms, and because calculating the expectation is not computable, AIXI cannot be implemented, and remains a theoretical model. The model is also flawed in that it completely separates the agent from its environment other than through the input/output channels, and so AIXI is unable to think about itself or change itself. But AIXI does provide a valuable abstraction for perfect intelligence. It provides the important insight that an abstract, powerful intelligence need have no built-in goals or human-like mental features whatsoever, other than being an optimizer.

When Artificial General Intelligence does eventually arise, it will probably not resemble any of these examples too closely. But understanding powerful optimizers like these helps us understand the many possibilities for intelligent entities.

Friendly AI

An Artificial General Intelligence that will not extinguish any possibility of a future which we would value is called a "Friendly Artificial Intelligence." (This is a term of art, and has nothing to do with friendliness in the ordinary sense.) In order to achieve Friendly AI we must create an initial AGI which fulfils two necessary conditions: (1) Its goals must be good for humanity and (2) it must retain its goal-system unchanged, as it self improves (See Bostrom & Yudkowsky, 2012).

We will have a valuable ally in preserving the AI's goal-system: The AI itself. Any change from the initial goals reduces the probability that the goals will be achieved, and

so a sufficiently powerful intelligence will fight against any change in its goals. (Humans sometimes change their goals, but we are fairly weak optimizers.)

Human values are complex, but at least they are stored in a known place: our brains. Yet more than understanding the complex and mutually contradictory desires of a single person, we must find a correct balanced combination of the distinct value systems of *all* human beings.

Research Today

Research into Friendly AI is in its formative stage. The two challenges in this research are defining the correct goal-system, and designing the intelligence so that it retains its goals as it self-improves. To accomplish this, researchers must combine insights from a variety of fields including decision theory, algorithmic information theory, formal logic, evolutionary psychology.

Today, the two leading research organizations in this area are the Future of Humanity Institute at Oxford University and the Singularity Institute in the Silicon Valley. Interest in the topic has been growing in academic circles, both because of the fascinating theoretical perspectives it provides, as well as the critical implications for the future of humanity. In 2011-12, the field grew rapidly: The noted philosopher David Chalmers published an article in the topic, and his colleagues responded with a dedicated issue of the *Journal of Consciousness Studies*. Researchers from the Future of Humanity Institute and the SIngularity Institute have published a number of peer-reviewed papers; and an edited volume, *The Singularity Hypothesis*, is scheduled for publication later in 2012.

Biography

Joshua Fox works in IBM, where he leads the development of software products. Before then, he worked as a software architect in several Israeli startups and growth companies. He holds a PhD in Semitic Philology from Harvard University and a BA *summa cum laude* in mathematics from Brandeis University. He is a Research Associate for the Singularity Institute. More information is avaliable at joshuafox.com

Further Reading

- An longer overview of Friendly AI for a general audience: Luke Muehlhauser, Facing the Singularity .
- An academic overview of Friendly AI issues: Nick Bostrom & Eliezer Yudkowsky, 2012, The Ethics of Artificial Intelligence To appear in T*he Cambridge Handbook of Artificial Intelligence*, eds. W. Ramsey and K. Frankish. Cambridge, UK: Cambridge University Press.