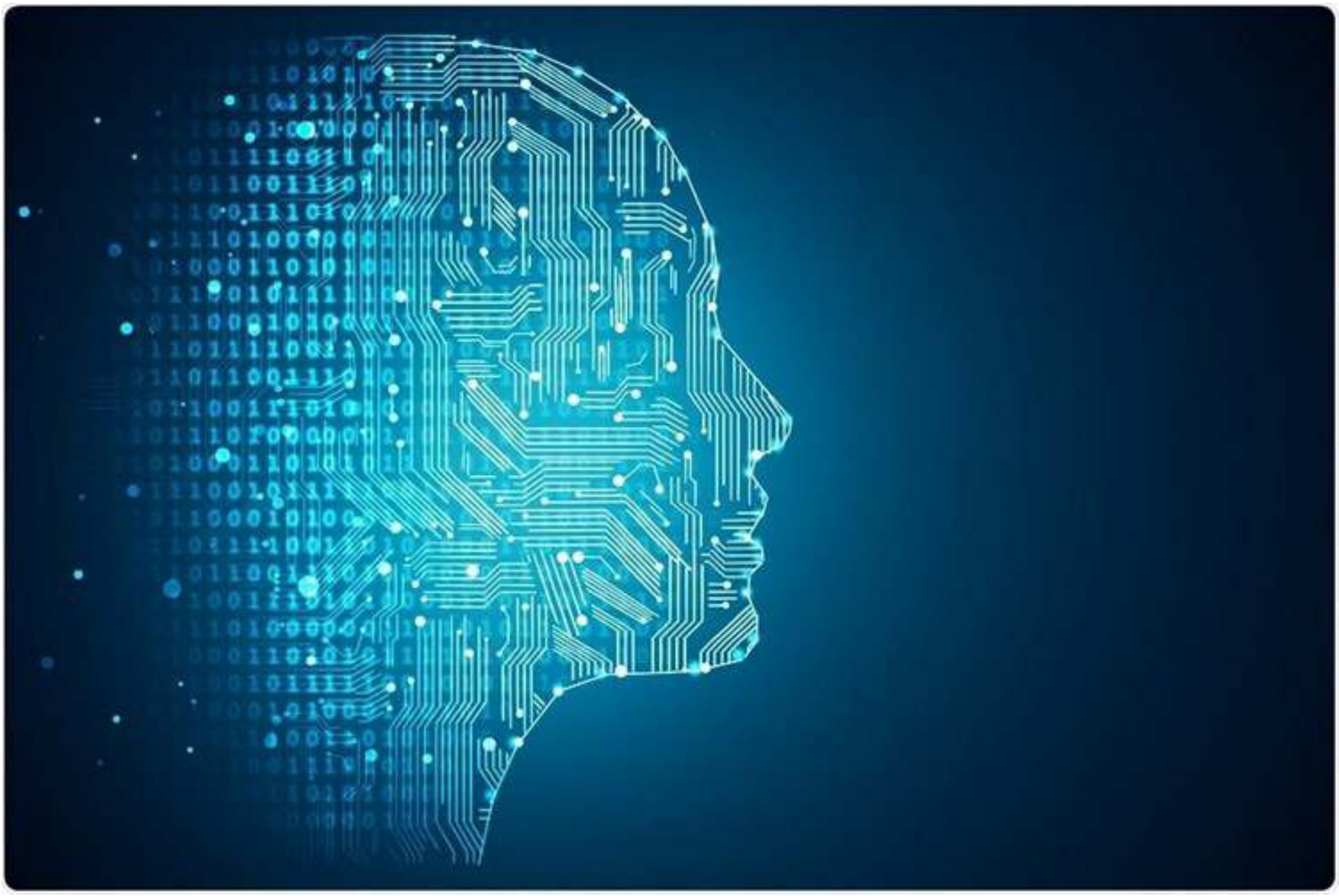
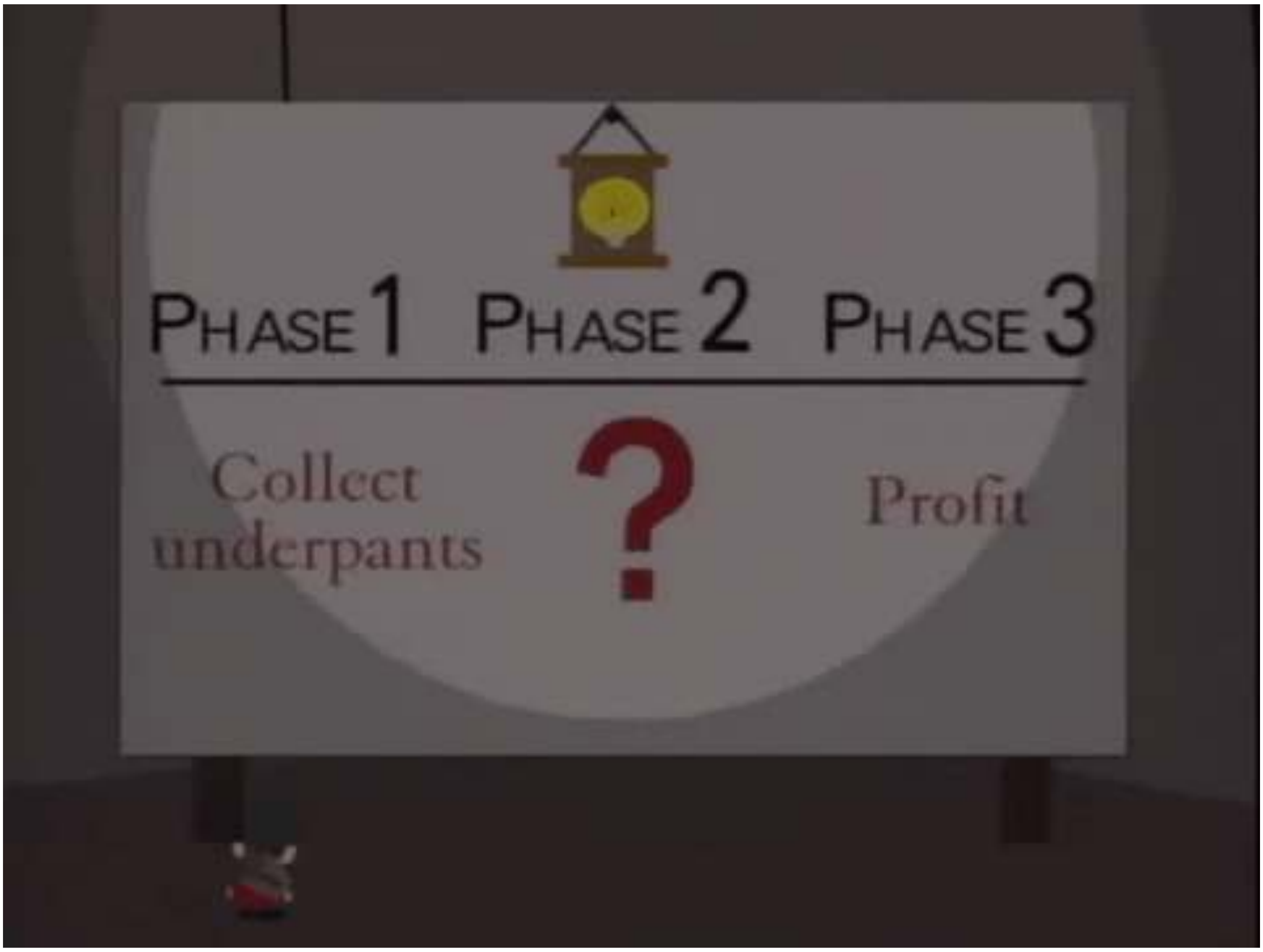
The New Frankenstein's Monster: Entering the Age of Artificial General Intelligence

Jeremy Owens



Credit: Peshkova/Shutterstock.com

I can tell you the business plans for the next 100,000 venture backed start ups in the coming decade: take existing thing, add AI, profit. Hopefully no one will be missing any underpants this time.

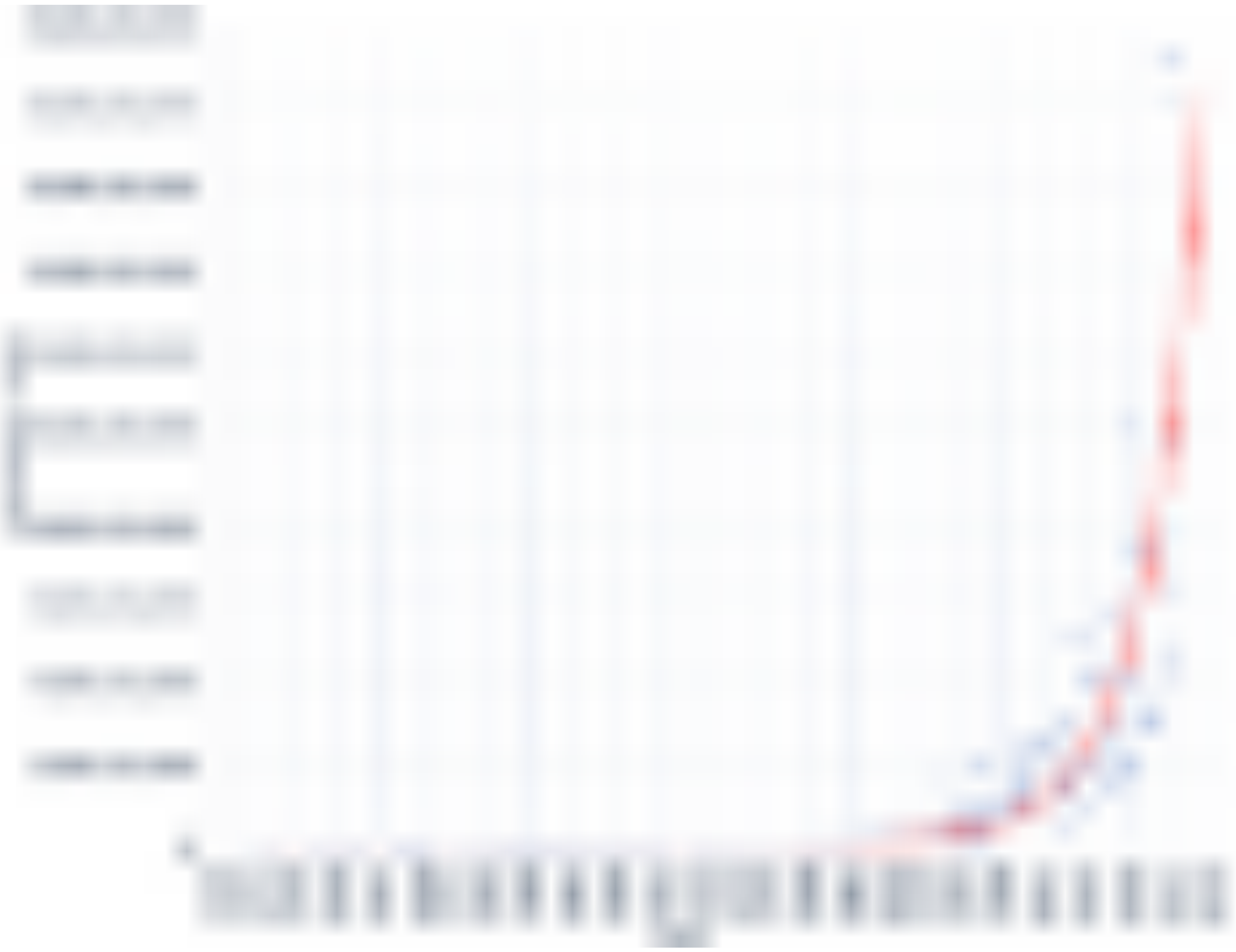

PHASE 1
PHASE 2
PHASE 3
Collect underpants
?
Profit

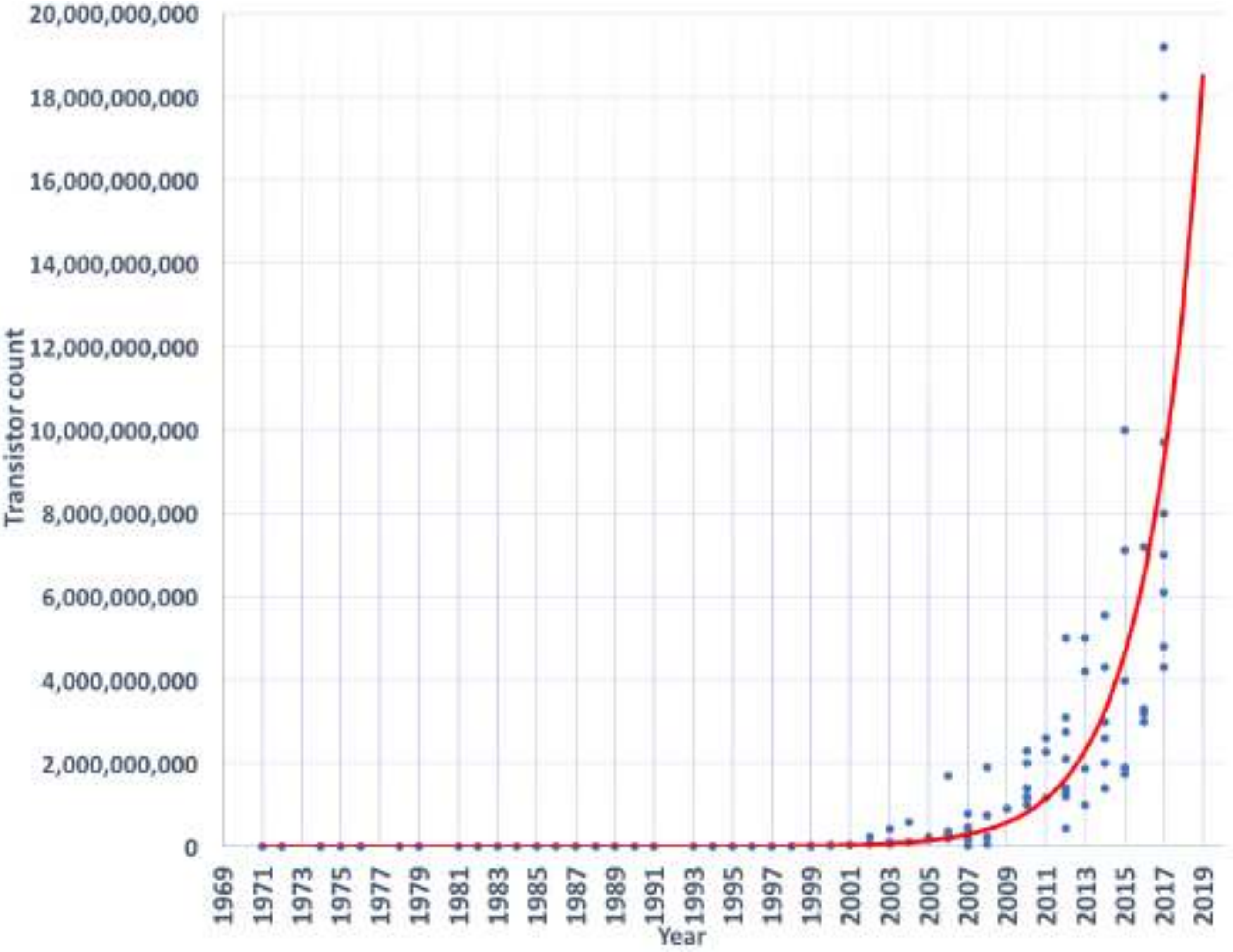
Pop culture references aside, the development of AI and machine learning has and will continue to dramatically improve the software we work with, shape our preferences, shift our interactions with other people, and expand our view of the cosmos. NPR, in a piece about how robots are the new space explorers, said people are relating to the robots sent into space the way we used to relate to astronauts, weeping when missions end and jumping for joy when Curiosity takes a selfie.

These robots we're sending to asteroids and the algorithms we use to tailor our news feeds and recommendations aren't our friends though, and they're not human. We're making tools, not colleagues. Until the ubiquity of the personal computer, almost all of the services humans enjoyed using came from the value that another human provided for us, so it's understandable that our language in reference to these services hasn't shifted dramatically. As our machine learning algorithms and AI learn more about us, learn more about our preferences, and can recommend to us what we'll enjoy the most, we have to shift away from an anthropomorphized view of these tools, because what comes next, after the AI of today, might well be something we interact with on a conscious level. Many researchers today call this next level Artificial General Intelligence.

Most of what we've created in this world of machine learning and AI (which I will refer to jointly from now on as AI) has been in the realm of Narrow AI. This means we have a specific problem that we want to solve, and we build an AI that can analyze and learn and optimize in that *one specific area*. This Narrow AI is what will power the development of businesses in the coming decades, though it's unlikely those businesses will create many new jobs. To apply that specific AI knowledge set to another area, even if closely related, is a massive undertaking though, and requires one to almost start from scratch. The most effective AI to work across disciplines is our favorite Jeopardy-stumping Watson, which is now used in the medical field for analyzing MRI scans for tumors and a myriad other medical analysis that can utilize big data. The problem is if you were to put Watson to the test to inspect a smoking engine and figure out what was wrong: it just wouldn't work. The ability to take knowledge from one area of life, transform it, and implement it in a totally new scenario never before seen is a uniquely human ability. The folks who are striving towards Artificial General Intelligence, or AGI, are trying to change that. They want to build something that resembles the kind of conscious thinking that humans experience, but why do we want to achieve this, and is it possible?

Can we, or even should we, be trying to work towards this kind of AGI? Let's first address the 'can we' question. From a purely computational perspective, the exact reconstruction of a human brain (if that's what we need for a conscious machine) is possible, it's just a matter of when. One of the most influential ideas in computational power is that of Moore's Law, that the speed and capability of computers doubles every two years.




20,000,000,000
18,000,000,000
16,000,000,000
14,000,000,000
12,000,000,000
10,000,000,000
8,000,000,000
6,000,000,000
4,000,000,000
2,000,000,000
0
Transistor count
1969
1971
1973
1975
1977
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
Year

Moore's Law past and projections

Some manufacturers are beginning to doubt that Moore's law will continue to hold, as the size of our transistors get closer and closer to the atomic level, where electrons begin jumping gaps when we don't want them to. The train hasn't stopped yet though, as Intel recently announced their chip stacking technology, to be released in 2019, that will continue this trend in computing power. Quantum computing is also on the way, and while its application won't be perfect for all purposes, for parallel processing, it will provide leaps and bounds more processing power.

On the more philosophical side of the 'can we' question, from a definitions point of view, Amir Husain, CEO of Spark Cognition, describes consciousness as such: "It's this self-awareness, this idea that I exist separate from everything else and that I can model myself…Human brains have a wonderful simulator. They can propose a course of action virtually, in their minds, and see how things play out. The ability to include yourself as an actor means you're running a computation on the idea of yourself." By this, we're remarkably close to development of AGI, as we already run algorithms to do this kind of thinking, though often without the machine itself involved. Were we to program a goal for the AI to improve its own computational ability and speed, would that not meet this definition set by Husain? Many people don't agree that this definition encompasses all of the conscious experience though, and we'll get to that a little later.

The 'should we' question is far more complex, but a lot of it will come down to business (or military) growth and development. If it can help a business do more, better, faster, or can save human lives, it's likely to arrive. There are a lot of arguments to be made that we should draw a line in the sand when it comes to AGI though. However autonomous AGI entities may become (which, considering the rate they could learn and adapt, is quite autonomous), they don't experience the vulnerability or mortality that we conscious humans share. Death is the great equalizer for all humans. If an AGI is able to save its data on a series of servers, and the current system is hit by a truck or 'blue screens', its memories and optimized processes can be brought back to a new corporeal being (should it even want one, really). It removes AGI from the world of the vulnerable and leaves them as de-facto immortal beings.

Whether or not we should work towards AGI, a diverse group of researchers and practitioners — computer scientists, engineers, zoologists, and social scientists, among others — is coming together to develop the field of "machine behavior,"…This field does not see robots merely as human-made objects, but as a new class of social actors. Even if we don't all agree that we'll arrive there, the problem at least has the beginnings of a solution. So what is the likelihood of these conscious beings arising after all? Whether conscious machines will or won't arrive has more to do with people's beliefs about human beings than their beliefs on the direction of AI. Looking at your own beliefs, what are we? What are humans?

There are a couple of options I'll lay out for you to consider: Machines, Animals, or Humans. Are we Machines, purely mechanistic beings, solely a product of the various chemical reactions, pheromones, and neuronal interactions we produce and respond to every day? Or are we Animals, something more than a machine, but something with an animating force that has thus eluded scientists (though not to say it forever will). If we're animals, can something mechanical get the spark of life? Or, are we Humans, and here I mean a distinction beyond the description of our Homo Sapien form. Are we more than just the apex predator of the animal world, are we unique in our views of our 'self' separate from our views of the other? Are those views our consciousness manifest, and is it unknowable to anything but humans? There are no right answers, but what one believes about humanity in general has a great deal to do with whether or not one thinks we can achieve AGI (machine yes, animal possibly, human unlikely).

What would a conscious machine even look like, and would we know if we saw it? Many people see consciousness as a binary option, either one has it or one doesn't, but consciousness may actually exist on a continuum. How one defines the limits of this continuum are still open for discussion, but there are a couple points to consider along this sliding scale. On the low end, the passing of the mirror test can serve as a rudimentary level of consciousness. If you put a living entity in front of a mirror, take the mirror away, put a dot on that entity's (equivalent) forehead, and put the mirror back, does the entity try to remove the dot? To do so would suggest that there is a persistent view the 'self' that entity has, that it recognizes has changed from seeing itself in the mirror between the first and second iterations. This experiment has some issues (define 'living'; requirement of vision; movement capable of removing the dot, etc.), but does at least give us a view into what beings, other than humans, might have a degree of consciousness about them. A level up on the consciousness ladder could be use of language. Some argue that nothing actually exists in the world until we give a word to it, so that others beyond our own interpretation of reality can share in what that thing is. That consciousness itself is a shared entity. It's possible that until you have language to determine self and other, there is no such distinction. Further up the ladder we have meta-cognition, or the ability to think about thinking. Humans who meditate do this on the regular, noticing their own thoughts pass by, but even non meditating humans have this ability, noticing when they're angry or sad, or in their analysis of another person's statistical proof. If you're willing to flex your consciousness muscles a bit more, you should also check out John Searle's Chinese Room thought experiment (often used to refute the possibility of AGI) that explores the difference between knowing and understanding something, the latter of which has been the sole domain of human beings so far.

The conversation around AGI has accelerated as people see AI getting smarter and smarter. Theoretically, this means we would reach a point where AI intelligence would be close to that of humans, and when AI reaches human level intelligence, it will likely also be conscious, or just like us. This is quite a jump though, as we have no idea how to begin the path to making AGI, and the path might just be superfluous. The path for our society may not require Artificial General Intelligence after all. Much of our work towards better AI is in an effort towards super-intelligence, knowing and having the capacity to know far more than we can possibly imagine in our current state. A conscious, intelligent being isn't the only option to reach that pinnacle of information, as Yuval Harari states in "Homo Deus", "There might be several alternative ways leading to super-intelligence, only some of which pass thought the straits of consciousness. For millions of years organic evolution has been slowly sailing along the conscious route. The evolution of inorganic computers may completely bypass these narrow straits, charting a different and much quicker course to super-intelligence."

For many of the problems we need to solve in the world, and for much of the exploration we want to embark on, while superior intelligence is a requirement, consciousness is not. Whether or not AI graduates into AGI may not actually matter in the end. There is a decoupling that has arisen from the development of AI between intelligence and consciousness. Where before we relied on humans to lead the path toward more a more stable, more advanced society, much of that work is now shifting to something we created, something that is beyond us, but something we already trust to do better than we could on our own.

There's a dearth of people working on this problem, of trying to solve the question of consciousness, what new forms of consciousness would look like, and how we make sure that what we create isn't something that leaves humans in the dust. 80000hours.org (a name based on the average number of hours a person works in their lifetime), consistently updates their problem profiles, and rates the scale, neglectedness, and solvability of various problems to solve in society. The positive shaping of the development of AI is rated as one of the most pressing problems to work on in our society right now: the risk of things getting truly out of hand are high, and it's a very neglected problem right now.

80000 Hours Assessment of the positive shaping of AI

80000 Hours Assessment of the Positive Development of AI

Thankfully, this means that individual effort in this field can have a tremendous impact. If this is an area that lights you up, I urge you to consider what you can do to contribute to this conversation and development. A great place to start is "The Fourth Age" by Byron Reese, which informed the machine, animal, human question above. It'll start you asking the questions you didn't know were a part of this debate. "Homo Deus" by Yuval Harari is a great next step, or for something with a bit more humor and illustrations, Tim Urban's 2 part post on super-intelligence.

As AI makes more informed decisions than we could ever accomplish, given its vast swaths of data that no human could ever conceivably know, what kind of control will we cede to it? Will it still be ours to manage, or will it be something entirely new we've created? It could get out of hand if we're not careful, and we need more informed people asking the right questions to shape how this future looks.