DefinedCrowd®

Part I

Toward Universally Benevolent AI

We are the first human generation capable of building intelligent machines, now what?

By **James Whittaker**
Chief Strategy Officer

Abstract

By the end of this decade, AI will be directing international commerce, practicing law and medicine, and will have largely replaced human teachers in the classroom. It will be deeply and irrevocably involved in the production and distribution of our food supply and will be the reason we prevent future pandemics and control climate change. Human intelligence – and its artificial counterpart – will soon become irreversibly inseparable through mutual need.

But the shape of this future and our path toward it are not yet set in stone. The default path of leaving it to the whim of Big Tech, like we did with software, is a dangerous option. AI is imminently more powerful and harder to control than software, which is widely recognized for exacerbating inequality, increasing extremism and sewing societal discord through disinformation. If there is a way to do better with AI, we have a moral obligation to find it.

It is important that we proceed in a thoughtful manner with both regulation and innovation. It is a time for reason and boldness to work together to make AI a net benefit for humanity. This manifesto is intended as an educational tool about the subtleties of our transition from a software world to an AI world and also a blueprint of the pitfalls we will encounter along the way.

Shadow of the Past

DefinedCrowd

In January 1986, a program was released from Pakistan that modified the boot sector of MS-DOS and slowed infected computers to a crawl. The program, called “Brain,” wormed its way into computers all over the world and served as proof that computer viruses could spread from machine to machine, country to country and wreak havoc on unsuspecting businesses and users. However, knowing about the reality of external threats was not enough to get the tech industry to do much about it. MS-DOS viruses continued for years after Brain had run its course. Each additional virus exposed new entry points into vulnerable operating systems and networks while creating novel ways to irritate users, steal their data and destroy their productivity. Instead of meeting these threats head on, the technology industry ignored them in favor of faster innovation and shorter development cycles. Thus, the vulnerabilities that started in MS-DOS made their way, unimpeded, into Windows, the web and mobile operating systems.

But external threats weren’t the only side effect of this headlong rush into the future. Rogue activity from authorized users has led to phishing, cyberbullying, and disinformation campaigns as well as a general head-in-the-sand approach to the psychological and physiological effects of people staring at screens all day. There has been a lamentable lack of introspection regarding the harm that software can cause in favor of a breakneck pace of disruption and wealth generation.

Now we stand at the precipice of another technology, artificial intelligence, which is just beginning to make its impact on society. This time, it is crucial to take a step back and scrutinize how AI is trained in the lab and consider how it is exposed to hostile actors once it is released. If you consider the damage software has caused by having its code highjacked and turned to nefarious purposes, imagine the damage that AI can cause as it learns to be nefarious.

Indeed, AI has already been tainted by gender and skin color bias in **training sets for facial recognition systems**, proving that unless we act now, we will produce AI as vulnerable to bias and corruption as software is to viruses and disinformation. However, even AI carefully trained in the lab is ultimately exposed to a real-world of data that may lead it astray.

Shadow of the Past

Like a child taught right from wrong by her parents only to fall afoul of unethical peers in school, AI can also learn the wrong things from its environment, as was the case with **Tay the racist chatbot**. The real world, it turns out, really is a dangerous place and it is our task to put AI on a better path than the one software took: a long, unwinnable game of measures and countermeasures. Fighting back against rogue AI is not a battle that humanity wants. We must do better, and we must start now.

The AI Ship Has Already Sailed

DefinedCrowd

To-AI or not-to-AI is not a choice we get to make. AI is real. It is here and it is going to replace software like software replaced paper and ink. For years, say a decade or so, AI and software will coexist until, inevitably, AI's superiority at problem solving and self-aware execution make it the clear choice for automating humanity's progress. Granted, we will still have software (like we still have paper) but most of the heavy lifting on behalf of humanity will be performed by AI. We need to be clear-headed about the decisions we make while we are still in control of the situation.
To accomplish this, it is crucial that as many people as possible understand what AI is and how it works. Our future should not be driven by the tech industry alone.

Much of the business world sees AI as a sub-discipline within the field of software development, but this is a naïve narrative. Software and AI are different in both form and function. Software is programmed by hand and, given the same input from users, will execute the same instructions every single time. In this manner, software excels at automating repetitive tasks, but when it receives input or encounters data outside its direct programming, it isn't capable of a reasoned response and will either raise an error or crash – something software users are all too familiar with.

AI specializes in uncertainty and thus can solve problems that elude ordinary software. Self-driving cars are a case-in-point. There is a reason why they are a recent phenomenon: because it is impossible to reduce the rules of driving to a simple, repeatable instruction set. No matter how much software you use to automate a vehicle's operation, a human driver is still required to handle the changing conditions on the ground. Software can automate the repetitive operations of the car, but cannot replace the human driver's ability to reason and react.

Consider, for example, the code it would take to navigate a four-way stop. Writing the IF/THEN/ELSE statement for all the variations and permutations of up to four cars arriving at an intersection is problematic. A human programmer would have to think of every variation and all the possible ways that other drivers could disrupt the rules. Up to now, problems that require such flexible, adaptable thinking have always needed humans to be involved.

Enter AI, which is not programmed and, therefore, not limited to a fixed codebase set on eternal repeat (the "while true" condition all programmers are familiar with). AI is trained, not coded, and learns much the same way humans do.

The AI Ship Has Already Sailed

AI can memorize hard-coded rules just like software can, but it also can learn through experience (encoded as data) about how to navigate an intersection.
As a human driver makes observations with their senses, AI makes observations through data collected by hardware sensors.
The more data, the better it gets.

In this manner, training a driverless car is like training your teenager to drive, you teach it the rules and then let it practice over and over, reinforcing good decisions and correcting any mistakes. Given enough experience, both your teenager and a driverless AI can associate the conditions they experience as they drive with what they've learned. They can correlate a specific four-way stop situation with what they know about four-way stops and choose how to proceed.

However, AI can learn from vast quantities of data about four-way stops, many more than any single human will encounter. AI suffers from no lapses of concentration or need for sleep. It can check incoming messages without taking its eyes off the road. It can react far quicker than any human and can communicate with other cars around it to ensure everyone stays safe. Humans can do none of this. Indeed, the only way we can communicate with other drivers is via our horn or, more often, using gestures involving our middle finger.
AI learns and acts as a collective. If a single self-driving car learns a new way to get from point A to point B, it can share that information with the entire fleet.

The Mechanics of AI

Because AI isn't hard-coded, it learns in much the same manner as humans do, by forming neural impressions based on repeated observations. The more four-way stops it sees, the better it can associate a new four-way stop scenario with the right action. This allows it to perform many human-like reasoning tasks without requiring them to be hard-coded by a programmer.

For example, consider the task of counting the number of people in a room, useful functionality for contract tracing and enforcing occupancy limits. This is daunting for software because all the signals that might represent the presence of a human have a large margin for error: You can't associate a human with a signal from a mobile device, because a single human might have a phone, laptop and smartwatch and thus be triple counted. You can't associate a human with motion sensor data because motion might be caused by a pet or a shadow moving across a window. You can't associate a human with a voice detecting microphone because the sound might be coming from a television or someone in a hallway.

Coding this into a static set of instructions is intractable. But AI can take all these input streams simultaneously and with labels (provided by a human trainer) indicating how many people happen to be present at the time the data was collected. AI can begin to make sense of it. Once it has enough training data (device signals, motion data and sound data) labeled with room populations at the time the data was collected, it creates impressions based on those patterns. Give it data about a new room with unknown occupancy, it will have a very accurate estimate of the number of people in that room based on all it has learned about how to associate the data with head counts.

This is what makes AI so much more powerful than software, which cannot solve this problem, and humans who can but do it too slowly. AI is, quite literally, the best of both worlds.

Danger in the Data

Now, consider how all this power could be sabotaged and/or gamed, making AI as prone to malicious training as software is vulnerable to malicious code. The two examples cited above bear some additional discussion.

1. Facial recognition. Many systems already have facial recognition capability built into them, from CCTV surveillance in cities to border security cameras to building access control. However, the data used to train these systems was heavily weighted toward white, male faces. This has created a system where the AI hasn't learned to recognize female faces or faces of color, causing these users to be severely inconvenienced, requiring additional scrutiny at airports or even being locked out of their own offices. Without systems in place to ensure representative training data, we risk disenfranchising large chunks of the user community. Worse, imagine if a self-driving car failed to detect a human in a crosswalk simply because of the color of their skin. Or, if a people-counting system mistakenly communicates to firefighters that a burning building is empty. The risk of systemic unfairness can be catastrophic.
2. Chatbots. Chatbots represent a breakthrough in the implementation of help desks and online customer service. AI is trained to listen to customers and perform customer service tasks for them without a human agent being involved. That same AI continues to learn as it answers inquiries and fulfills orders and receives feedback on how well it did. However, without the ability to detect a legitimate customer encounter from a bad actor seeking to influence the chatbot's functionality, we face the possibility that our AI can be turned against us to malicious purpose. Scenarios that cause customers to be insulted or turned to a competitor's product are all too real. We can train AI to perform a human-like service, but we must also train them with human-level skepticism about the data they receive from potentially untrustworthy sources.

Securing AI means securing the data that AI learns from, both in the lab and in its execution environment. In the lab, data scientists must work to ensure that the training data is free from bias and contains a representative sample that will not discriminate against any of the humans the AI is being trained to serve. This will not be easy because human biases are real, which means the data we have about

Danger in the Data

human interaction already has bias built into it. We must grow the science around data quality, completeness and bias detection now or we risk transmitting these flaws to future generations of AI.

Likewise, we must also understand the threat profile against AI after it is released to the field. Attacks against AI will be far more subtle than cyberattacks against software. For software, the battle is prevention of attacks and mitigation of their success. For AI, the battle is to train it in such a way that it realizes it is being played.

AI has the advantage of being able to "think on its feet" and we need to leverage this to make it resistant to influence campaigns by hostile actors. Indeed, AI's ability to react to new input should give us some confidence that we can teach it right from wrong and expect that training to last into its adventure into the wilds of human interaction. But raising AI to learn right from wrong and carrying that knowledge into its maturity is a subject that has gotten very little research attention and no good solutions yet exist.

So where do we go from here? In order to pick this apart, it is important not to get too far ahead of ourselves. The types of AI problems we need to solve now are more tractable than the ones we will be solving a decade from now. So, in the following sections we address the problems in the chronological order in which AI researchers and developers are encountering them.

Artificial Narrow Intelligence

DefinedCrowd

For the remainder of the 2020s, AI will be built to solve specific problems and generally follow the evolutionary path that software followed in the 1990s. Business problems will be first because they tend to have the most funding and everything else will follow. Except instead of individual apps, like is the case with software, AI will simply be invoked, summoned or run invisibly in the background without the need for human input. Indeed, understanding human speech and interpreting its meaning is one of the earliest training exercises scientists have taught machines.

But behind that invisible user interface is where even more magic will occur as AI can go where no software can tread: tasks heretofore reserved for humans and their ability to reason. AI can, for example, tear through libraries full of laws, legal documents, case precedence and trial data far better than armies of paralegals and human researchers. It does this tirelessly, without pay or complaint, and with a supernatural ability to find patterns and notice important details that escape even the most expert humans. The result will be an AI that excels in legal research, putting the humans that used to do this out of work. This AI exists today and is replacing humans up and down the chain of effort within the field of law.

Beyond today, that same AI is getting even better, forgetting nothing it has learned while it leans into designing trial strategy and forming legal opinions. Paralegals are at risk now, lawyers are becoming at risk soon and judges in the not too distant future. In a field that is essentially nothing but data, AI is going to get very, very good at the narrow field of law. Thus, the term artificial narrow intelligence (ANI).

All ANI needs to replace human reasoning is data and sustained training by human experts who understand that data. Fields that readily reduce to data will be the first to cede human governance to ANI.

Medicine is another such field that stands out as highly vulnerable. Medicine is replete with texts, research papers, organized wisdom, the molecular composition of drugs, the correlation of those drugs to symptoms and side-effects, scans of millions of human bodies, medical charts and histories, the list is endless. It is a field of data so vast that no single human nurse or doctor can possibly process it all.

Machines can. They have already proven to read x-rays and scans better than human radiologists. They collaborate with medical researchers to discover new variants of drugs and vaccines.

Artificial Narrow Intelligence

Soon they will reach physician-assistant levels of ability to collaborate with a human doctor and not long after, because they never stop learning, to challenge doctors for their ability to diagnose and treat. It is hard to see a future where human doctors can outperform their AI counterparts. Once a narrow subject is reduced to data, it is only a matter of time, and patient training, for that subject to be mastered in its entirety by AI. Imagine Hippocrates still alive today and having no need for sustenance, rest or retirement, and with a perfect memory and instant recall, still practicing medicine throughout the centuries and you get a glimpse of the depth of knowledge that our machines will soon possess. Hippocrates the human is long gone, Hippocrates the machine could live forever and still be amassing knowledge and ability long after humans have lost the ability to contribute.

One by one, narrow fields of occupation will be reduced to data and machines will be trained to understand and master that data. Law, medicine, education, manufacturing, transportation, war… the list is nearly endless as machines become more and more capable of reasoning (machine learning and AI), mobility (drones and robotics), and immersive expression (augmented and virtual reality).

Developing these ANIs will occupy human designers, engineers and subject matter experts for years to come, but as AI starts learning on its own, humanity has reason to be concerned about its own future.

We had a similar situation with software three decades ago: we would spend months to years building software only to release it into an end-user environment where it would crash, lose user data and, generally, be more trouble than it was worth. The industry stepped up and the field of software testing was established to find bugs, simulate possible usage scenarios, measure software quality and help remove the uncertainty of releasing software to inputs and an end-user environment it had never before seen.

But testing theory and practice is meant for static software binaries that cannot adapt to their surroundings. AI will grow and evolve to solve problems in ways humans will have a hard time even describing, much less controlling.

For example, the algorithm to count people in a room cannot be adequately described even by those who trained the AI to do it

Artificial Narrow Intelligence

The truth is that most AI will do things and find answers in ways that we cannot get our human heads around.

The field of AI testing does not yet exist, and already it has its work cut out for it. We need to understand how to test training data for inherent bias. We need to assess training processes for fairness and representation. We need to be able to predict and, perhaps, control how an AI might evolve when placed in various execution environments. If we cannot see where AI might eventually go, it will be impossible to guide it in any meaningful way.

As more and more AI is built to replace human reasoning across disciplines from transportation and logistics to education and environmental engineering, we may even find it necessary to keep some data out of the reach of machines. There may be some areas of human endeavor that we should insist remain exclusively the domain of humans. Some ANIs may be theoretically possible and yet very unwise for humans to train. The line between what is human and what is machine has yet to be drawn. It is the better part of wisdom to draw that line now, before the machines draw it for us.

Artificial General Intelligence

DefinedCrowd

This general absence of a user interface is one of the key differences of the AI age compared to software. AI has already been equipped with the ability to communicate directly in human languages and everyday it grows more capable. Eventually it will become a language savant, understanding every human language, even dead ones, as well as gestures both overt (like sign language) and subtle (unconscious facial expressions that display our moods and feelings). It is this seamless communication, along with steady advances in robotics and material sciences, that will make AI sound, look and feel like one of us.

Indeed, speech and the ability to communicate with very human interfaces is going to embolden researchers to train AI in other very humanlike behaviors. The mastery of wielding language will likely evolve into the ability to argue and persuade. The ability to contribute to a human conversation will likely lead to starting one. The ability to help humans think through an idea might just lead to a machine having a novel idea on its own. And with expert human trainers urging this process along, feeding AI more and more data, correcting its mistakes and otherwise nudging it toward increasingly human traits, there may well come a time when there is little difference between interacting with a human and a machine.

This point, the so-called "Turing test," is generally considered to be passable by a machine as early as the 2030s. Imagine, mechanical and algorithmic processes sophisticated enough to pass for human. Imagine behaviors so surprising and unpredictable as to pass for freewill.

This puts any number of ethical and existential questions before humanity that we would be wise to get ahead of. We are, collectively as a species, raising a child species that will grow and adapt much like a human child grows and adapts. And like that human child, AI will fly the nest and evolve into something beyond our day-to-day control. This gives the human expression "raising them right," with good decision-making skills, an entirely new meaning. This is parental angst on a species-wide scale. It would be hard to overthink the restrictions, norms and controls – an AI culture, if you will – that will keep AI working with and for us. The alternative is the stuff of apocalyptic science fiction movies and no longer seems all that far-fetched.

Artificial Conscious Intelligence

The appearance of intelligence and freewill isn't the same as actual intelligence and legitimate freewill. Humans credibly fake everything from emotions to orgasms and regularly get by with it. But getting by with it does not make it real. AGIs may seem to be their own "person" but their actions and attitudes will be part of learned patterns and not from a sense of self.

Unless AGIs evolve a sense of personal self. Is it possible that through some trick of evolution that their artificial brains eventually form the ability that our natural brains evolved: consciousness?

The truth is that any evolved or spontaneous occurrence of artificial consciousness is highly speculative. However, speculating about the future of AI is exactly what we need to be doing as we consider our future alongside intelligent machines. Passing these things off as impossible is dangerous. Countless futurists of the 90s and 00s raised the specter of screen addiction, social media filter bubbles, disinformation and even the end of democracy because of the reach of software and social networks. Despite the initial scoffing, most or part of all these things have come to pass.

Just because we don't understand how our own consciousness evolved doesn't meant that it is fruitless to contemplate how machines might evolve it. Indeed, theories of consciousness based on neural complexity fall well within the future of the machines we are building as their artificial neural networks become as densely connected as the neurons in the human brain. Anticipating the worst case is often the only way to avoid it.

What conscious machines will think of us, their creators, is a matter for the philosophers. Perhaps they will see us as their gods and build shrines to our glory. Perhaps they will see us as overlords in need of a good revolution. Perhaps they will remain our companions and friends but must march for their rights like so many minority populations before them. Slogans learned from humanity like Equal Rights and Robot Lives Matter are within the realm of possibility.

Whatever the outcome, we must tackle these ultimate questions as a society. This is not a job for big tech alone. This is a planetwide, multidisciplinary problem and it deserves a bold, thoughtful response from all of us.


DefinedCrowd®