WHITE PAPER

THE STATE OF ARTIFICIAL INTELLIGENCE

T-Systems in collaboration with
the House of Beautiful Business

Written by Andrian Kreye

November 2018

Based on interviews with leading artificial intelligence (AI) thinkers and practitioners, this white paper focuses on the current opportunities of narrow and broad AI, debunks the grand myths of AI, lays out the debates that are currently shaping the approaches to developing this powerful technology, and shows how both society and the technology community can learn—and benefit—from the mistakes of 20 years of digitalization.

INTERVIEWEES

John Cohn
IBM Fellow, Internet of Things Division

John C. Havens
Executive Director, The IEEE Global Initiative on
Ethics of Autonomous and Intelligent Systems

Sven Krüger
Chief Marketing Officer, T-Systems

Jaron Lanier
Computer Scientist, Author, and Composer

Peter Lorenz
Senior Vice President Digital Solutions, T-Systems

Manuela Mackert
Chief Compliance Officer, Deutsche Telekom

John Markoff
Fellow, Center for Advanced Study in the Behavioral Sciences

Kenichiro Mogi
Brain Scientist and Senior Researcher,
Sony Computer Science Laboratories

Radhika Nagpal
Professor of Computer Science, SEAS and Wyss Institute,
Harvard University; Co-Founder and Scientific Advisor, Root Robotics

Steven Pinker
Johnstone Family Professor, Department of Psychology,
Harvard University

Tess Posner
CEO, AI4ALL

Mark Rolston
Co-Founder and Chief Creative Officer, argodesign

Sarah Spiekermann
Chair, Institute for Management Information Systems,
Vienna University of Economics and Business

Jaan Tallinn
Founder, Future of Life Institute and Co-Founder,
Skype and Kazaa

Max Tegmark
Founder, Future of Life Institute and Professor
of Physics, MIT

Ulli Waltinger
Founder and Co-Head, Siemens Artificial Intelligence Lab

HAVE YOU EVER MET AN INTELLIGENT ROBOT?

There aren't too many around yet. Most robots still busy themselves with menial tasks like vacuuming apartments or placing components into cars on assembly lines. But if you do meet a machine equipped with AI, it can be a daunting experience. Take Spot. Constructed by Boston Dynamics, Spot is a dog-like machine that can easily traverse any terrain, possesses great physical strength, and incorporates an AI system that allows it to move about independently.

Once in a while, Spot gets to leave its company labs and testing grounds to meet regular people. On a recent sunny afternoon, it stood in the lobby of a large conference center. A machine on four legs, about a meter high with a weight of 75 kilograms, it crouched as if ready to jump, turned, and took a few steps towards the onlookers. Anyone who touched it could feel the vibrations of its electric motors. In this incarnation its engineer didn't even try to make it cute, as they generally do with robots prepped for tech-demos. Spot has an athletic body of steel and hydraulics, the components and cables of its guts clearly visible, an appearance more attack dog than benign companion.

Although the menacing design might be a marketing gimmick geared toward potential military buyers, Spot will be mostly deployed in disaster relief and private industry. It does not yet have great intelligence, but it does have enough smarts to work independent of human direction. It vividly illustrates the two sides of the AI debate: Will this smart machine be pit bull or St. Bernard?

Right now, artificial intelligence is defined in three tiers. Narrow AI is any program that is driven by self-learning systems of algorithms. That includes the customer preference service of shopping platforms such as Amazon or Otto as well as the feeds of social media platforms such as Facebook or Xing.

Broad AI is any machine-learning system that can optimize itself. These systems do so either through the analysis of big data (such as most self-driving cars) or through deep-learning methods such as Google's AlphaGo program, which has learned to play the board game Go.

Third tier systems can learn by themselves when given a set of rules. AlphaGo's successor, AlphaGo Zero, is arguably the most famous of these: it quickly surpassed all known playing levels by playing against itself. Also known as Artificial General Intelligence (AGI) or "strong AI," these systems could combine a wide variety of skills, achieving human-like intelligence, and—according to some academics—a degree of consciousness.

This last possibility is still a mere concept. Estimates of when AI technology might develop to the point of consciousness range from five or ten years to never. But one thing is certain: With the continued development of AI, technology, society, and everyday life will change profoundly—and Spot is the perfect embodiment of that promise.

PART 1: THE INFLECTION POINT

What makes AI so exciting right now is that we have reached a multilayered inflection point, a moment when we have the opportunity to learn from past mistakes and deliberately choose how to guide AI's trajectory into the future.

For the first time since humans discovered levers, the relationship between man and tool is about to change. Although prehistoric people used branches to move a rock—and people of the digital age activate an app with the swipe of a finger—the effect is the same. People initiate action. People are in control. With self-learning machines, decision-making processes and actions will become independent of human interaction.

> *Mark Rolston, founder of argodesign in Austin, Texas, and creator of the first modular system for a mass market AI:* ***"We're starting to develop a relationship to machines where machines are less like tools, like a hammer, and more like domestic animals, like a dog."***

But it's not surprising that this watershed moment generally is not perceived as such. Technological progress has been seeping into life at a pace that rarely concerns most users. Low-level AI is already part of daily life, be it through digital assistants such as Siri and Alexa, driver assistance features in cars, or the combination of self-learning algorithms and the hive mind of Facebook and Twitter user bases. Advances in digital media, genetic engineering, and weak forms of AI are already taken for granted. Few technological leaps seem to astound a general public inured by a decades-long barrage of new smartphones, social networks, and digital assistants.

> *Sven Krüger, chief marketing officer of T-Systems International:* ***"John McCarthy, one of the pioneers of AI, was right in his observation that as soon as it becomes an everyday technology, it is not perceived as AI anymore. The algorithm sorting search results for example—now we say, we google something. Which is a good thing. … AI is merely a tool of digitalization. It will become as natural as electricity, because it will be part of almost everything in which electricity flows."***

At the same time, the downsides of the ongoing digital revolution have become glaringly obvious. This has fueled a small but significant retreat from technology, mostly by educated elites in Western societies, and a more widespread desire to engage in thoughtful and critical conversations about where we're headed.

A healthy skepticism has begun to steer developers, entrepreneurs, and lawmakers away from the euphoric belief in a digital utopia that has already resulted in so many unintended consequences and missteps, including:

- The rise of a digital monopoly dominated by the Big 6: Amazon, Apple, Facebook, Google, IBM, and Microsoft
- The disappearance of privacy and media sovereignty, possibly enabling surveillance societies
- The behavior modifications deliberately induced by technologies such as smartphone game apps, Facebook, or Snapchat

- The erosion of public discourse, witnessed most prominently in the US and Europe
- The damage to the democratic process, which resulted in the rise of populist leaders, parties, and movements such as Donald Trump in the U.S., the AfD in Germany, and the Brexit movement in the UK
- The abuse of digital technology by authoritarian and criminal forces, most prominently by Russia, China, North Korea, and black hat hackers located in those countries, as well as by terrorist organizations such as the IS

> *Jaron Lanier, a pioneer in the field of virtual reality, one of the most prolific insider critics of Silicon Valley, and author of the bestselling book* 10 Arguments for Deleting Your Social Media Accounts Right Now*, says:* ***"I don't believe our species can survive unless we fix this. We cannot have a society in which if two people wish to communicate, the only way that can happen is if it's financed by a third person who wishes to manipulate them."***

If these mistakes are treated as lessons, the rise of AI could become one of the most exciting and transformative developments in technological history.

PART 2: THE STATE OF THE DEBATE

The most difficult part of the current debate about AI is the cognitive dissonance between current technological reality and the imponderables of a technology developing both in unexpected bursts of progress and unbelievably slow increments. That dissonance has led to a polarization between two extreme views which have dominated the public discourse.

On the one hand, AI enthusiasts hold a utopic vision of a future in which AI merges with and overtakes humankind, a hypothetical event called "technological singularity." First used by mathematician Vernor Vinge in 1983[1,2], the term has been most successfully popularized by Ray Kurzweil and his radical vision of creating a better human being. The author of several bestselling books, Kurzweil also founded Singularity University, a think tank with educational outreach programs.

In the spring of 2018, Kurzweil laid out his vision as follows: "In the 2030s, we will merge with the intelligent technology we're creating. Two million years ago we got this additional neocortex and put it at the top of the hierarchy, and that enabled us to invent language and technology and conferences, something that no other species does. Now we're going to create synthetic neocortex in the cloud. And just as your phone makes itself a million times more capable by connecting to the cloud, we will connect the top layers of our neocortex to the synthetic neocortex in the cloud. And just like two million years ago, we'll put that additional neocortex at the top of the neocortical hierarchy. Only this time it won't be a one-shot deal. Two million years ago, if our skulls had kept expanding, birth would have become impossible. But the power of the cloud is not limited by a fixed enclosure. It's doubling in power every year now as we speak. So we will have an indefinite expansion of our neocortex, and just like two million years ago, we will create new forms of expression that we can't even imagine today. If you then do the math, we will expand our intelligence a billion-fold by 2045. That's such a profound transformation that we borrowed this metaphor from physics and called it a singularity."

Those at the opposite end of the spectrum fear the future will bring runaway AIs that destroy humankind in a hyperrational, rather than malicious, way. The canonical thought experiment for this apocalyptic view of AI's future is the parable of the paperclip maximizer, as first described by the Swedish philosopher Nick Bolstrom in 2003[3].

Bostrom's hypothesis is based on an AGI, a general artificial intelligence, that it is not specialized in a narrow set of tasks like current AI, but capable of combining a vast array of activities and thought models much as humans do. In Bostrom's thought experiment, this AGI is tasked with maximizing the number of paperclips it owns ad infinitum. To fulfill its mandate, it will expand its own intelligence to the point of an intelligence explosion.

After surpassing humankind in skills, as the story goes, this AGI will use all means necessary to create ever more paperclips, depleting the planet of resources, destroying the environment and, as a consequence, humankind itself, ultimately creating paperclip manufacturing facilities in space, thus expanding indefinitely, destroying everything in its path to achieve its goal.

In 2014, Bostrom expanded on his ideas and thought experiments regarding the dangers of AI in his book *Superintelligence*. Two of his most prominent readers were physicist Stephen Hawking and Tesla inventor and space explorer Elon Musk. Both Hawking and Musk began to speak out publicly, warning of the imminent dangers that AI superintelligence posed for humankind.

What started as a philosophical thought experiment has now warped the debate about AI to a point bordering paranoia. However, at least one person—Jaan Tallinn, one of the founders of Skype—has been able to constructively channel the passion surrounding the topic. Tallinn founded both the Future of Life Institute in Cambridge, Massachusetts, and the Centre for the Study of Existential Risk in Cambridge, England, to try to influence the direction of AI development. The two organizations he founded share many well-respected and prominent board members, including Elon Musk and Nick Bostrom, who are committed to investing in a concerted effort to create safe and benevolent AI.

> *Jaan Tallinn sees the dangers that AI poses less as a result of a maliciousness and more as a consequence of humans not realizing the powers of the machines they create:* ***"As long as we are talking about things that are dumber than humans, we can treat it as just another technology. Things might go wrong. It might create accidents, it might create unemployment. Even if we get it right, AI might have big societal effects. Once we have something that is smarter than humans, the situation does change. There are many metaphors we can use. They are never precise. But look for example at the gorillas. They become extinct because humans have destroyed their environment. Not out of malice or hatred of gorillas, but because of disinterest in their fate. In some ways, the fate of the planet is always in the hands of the smartest agent. So far this has been the human. The concerns are about something turning really competent while being indifferent to humans."***

There are severe weaknesses in the arguments on both sides of the debate on the future of AI. The paperclip and other metaphors popularized by AI skeptics are largely based on the idea of a single objective AI igniting a powerful chain reaction of highly rational causes and effects. This is unlikely. (It should be noted that AI critics also point out some of the more imminent dangers fueling their concerns, which include the use of AI for military purposes, and the automation of millions of jobs, which could create a long period of mass unemployment, poverty, and civil unrest.)

The singularity vision, on the other hand, assumes that a machine will use reason and logic to process humongous amounts of data, while at the same time using robotics which would allow it to become independent of human interference[4]. Eventually, this would enable it to produce not just software (which is very realistic), but also self-designed hardware (which may eventually be possible). What singularity will not achieve, according to most experts, is turning AI into a sentient being.

> *Kenichiro Mogi, neuroscientist and founder of the Sony Qualia Lab:* ***"We will not be able to create artificial consciousness. AI is based on statistical learning. Qualia, the subjective, personal experiences that are at the core of our consciousness, cannot be built from a statistical approach. If you look for consciousness in AI, the right analogy is dropping a key in the dark and looking for it under the streetlamp. You look in the wrong place, because you think it is there, but it isn't."***

While a conscious AI remains in the realm of science fiction, the benefits of AI have been vastly underestimated. The main benefit AI provides now and will continue to provide is the ability to analyze amounts of data impossible for humans to do themselves. Technologically, the current moment is a perfect storm of developments. The advances made in processing power, big data storage, and robotics over the past few years have made things possible now that were distant dreams for most of the short history of AI development. In science and medicine this could bring about advances not yet imagined.

Consider these coming realities: the development of artificial synapses, which will allow the neural networks of AI to connect in ever more complex ways and exponentially reduce the amount of energy we use; the emergence of stable quantum chips, which will open new possibilities for security and processing power; and the first signs that AI will master—and not just process—human speech.

The biggest leap will be the merging of AI with robotics. This will create an Internet of Things capable of much more than the fridge ordering more milk or the networked thermostat adjusting room temperatures. For example, Radhika Nagpal, a roboticist at Harvard University, has programmed cheap mini-robots that fit into the palm of a hand to collaborate without human interference. A thousand scattered insect-like machines can gather and move amongst each other, following their programmed rules of engagement, until they have formed, for example, the letter "K" or the shape of a starfish. Nagpal believes this new form of collective intelligence—modelled not on human intelligence, but on the kinds of intelligence found in nature—will lead to new possibilities for collective behaviors.

> *Radhika Nagpal, roboticist at Harvard University:* ***"There are no tight definitions of AI and robotics anymore. Take self-driving cars. Is that AI? Or robotics? I think it's both."***

All the women and men currently involved in the creation of AI are not simply writing code and creating new technologies. AI is not just a mechanism, it is a highly developed new entity which will be learning tasks that have so far only been attributed to the human brain. Today's engineers are writing the DNA of AI. They are laying the foundations of a new set of values and a new definition of the relationship between human and machine.

PART 3: NEW VALUES

The AI debates have helped spur the widespread recognition that current laws and general rules of ethics are inadequate to address what's coming. For society, lawmakers, and developers it will be important to balance the debate in a rational way. This remains difficult as long as there is no real understanding of this new technology, and no clear set of common values applied to its development. Both needs are slowly being addressed, so far mostly by research institutions and companies. No one wants a repeat or expansion of the mistakes that led to the abuses of privacy, autonomy, security, justice, and power outlined earlier. Yet in the rush towards a digitized future, lawmakers, governments, and societies have often appeared to lag behind.

Whistleblower Edward Snowden's exposure of a worldwide surveillance program by the US National Security Agency—and its partner agencies of the "Five Eyes"—in the summer of 2013 was arguably the single biggest cause of a darkening global mood concerning digital technology. The second big event was the revelation of the spate of digital breaches during the US elections of 2016 —the hacking of the Democrats' email server, the leak of emails damaging Hillary Clinton days before the election, the exposure of massive manipulation of social media by Russian agents, and the business practices of Cambridge Analytica.

The third factor was the turnaround of Silicon Valley insiders like Tristan Harris of Google, Sean Parker of Facebook, and Microsoft's Jaron Lanier. They revealed the existence and pervasiveness of behavior modification mechanisms geared to promote the addictive use of digital technology. These mechanisms had been deliberately built into social media platforms such as Facebook, and the user interfaces of many smartphones.

While there have been some sensational consequences of the ensuing techlash—the five billion dollars the European Union fined Google for antitrust violations, for example, and the drop in Facebook stocks after both revenue and customer base started to shrink—one positive outcome has been the move to create values and ethics frameworks for the development of AI. In 2015, Elon Musk started the nonprofit organization Open AI as a think tank and research group devoted to developing safe AGI supported by Amazon and Microsoft. Amazon, Facebook, Google, IBM, and Microsoft formed the Partnership on Artificial Intelligence to Benefit People and Society in 2016 to set industry standards and best practices.

Europe, and especially Germany, play an important role in this creation of new values. Deutsche Telekom was one of the first global companies to draft, adapt, and publicize a code of ethics for AI. When the nine-part guidelines were published in May of 2018, they were a groundbreaking statement for an industry still looking to find its ethical bearings. It also led to Deutsche Telekom joining the Partnership on Artificial Intelligence to Benefit People and Society as one of the first non-American organizations in June of 2018.

> *Manuela Mackert, Chief Compliance Officer, Deutsche Telekom: "**More and more products and applications are driven by artificial intelligence. We do want to use the possibilities and potentials of AI for humankind, but with an ethical framework based both on our values of self-determination and our common sense. That is why it was important for us to find ethical guidelines which can directly affect the work of developers, programmers, and engineers. We worked with an interdisciplinary team and reached out to large number of companies and organizations to ensure our guidelines will be state of the art. We are now actively implementing these guidelines. We distribute educational material to our staff, we organize workshops for our clients, our internal processes are based on digital ethics, and we are currently forming internal committees to oversee the development of our products and services. But we do not just want to establish frameworks for our company, but initiate a public debate with citizens, politicians, consumer organizations, and other companies. "***

Germany in particular, with its first wave of laws such as the Network Enforcement Act (holding internet entities responsible for hateful content), or the General Data Protection Regulation[5], ensuring basic privacy for users—both enacted in 2018—is in the vanguard of this movement by not only defining, but enforcing a canon of values.

> *Peter Lorenz, Senior Vice President, T-Systems Global Systems Integration: "**We've seen incredible developments in AI over the last five years. Ten years from now, billions of AI-based devices will be in use. As powerful as they have already become, machine-learning systems are not the equal of us, because our intelligence is of a different kind: Humans also learn on***

> ***a meta-level. We explain what we are doing. We think about how we are thinking. And most importantly, we listen to, learn from, tell, and compose stories, and that still separates us from machines and probably will forever. However, AI doesn't know any ethical or moral boundaries. Imagine the consequences of AI-based bots adopting or absorbing human prejudices. That possibility demands or requires that we assume a new digital responsibility.***

The most fundamental work on AI ethics, however, is still being done in the academic domain. In January 2017, the Future of Life Institute hosted the Beneficial AI conference at the Asilomar Conference[6] grounds in California. The organizers deliberately chose the location to evoke another pivotal meeting held there in 1975, when the Asilomar Conference on Recombinant DNA established the guidelines for biotechnology. Those guidelines ensured that research and development would be conducted to benefit humankind and society, a goal so far achieved on a broad basis. The meeting and the publication of the guidelines also launched a groundbreaking public discussion of science policy, which ultimately led to a moratorium on certain experiments and avenues of research.

The 2017 Beneficial AI conference resulted in a declaration of 23 guidelines called the Asilomar AI Principles[7]. The principles covered research parameters, safety concerns, and longer-term issues, culminating in the final principle which states: "Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization."

A who's who of over 1,200 AI researchers and developers including Demis Hassabis, Ray Kurzweil, Francesca Rossi, and Stuart Russell endorsed the paper as well as more than 2,500 other people including Elon Musk, Erik Brynjolfsson, Maria Spiropulu, Christine Mitchell, and the late Stephen Hawking.

> *Max Tegmark, physicist and co-founder of the Future of Life Institute:* ***"We should get into the habit, when we build machines today, to make sure that their values are aligned with ours. So as they get more advanced, we are ready to take the next step. Everything we can do now is stepping stones to what needs to happen later. Asilomar in the 1970s led to a moratorium for certain kinds of biology work. We put our conference there for symbolic reasons. One of the Asilomar principles says we should avoid an arms race on lethal autonomous weapons. So most AI researchers would very much like to see a moratorium on slaughter bots and that kind of stuff."***

Another new potentially game-changing organization is the Council on Extended Intelligence (CXI). A joint effort between the IEEE Standards Association and the MIT Media Lab, the Council's members include luminaries such as the head of the MIT Media Lab, Joichi Ito, Harvard law professor and former presidential candidate Lawrence Lessig, and Columbia University economist Jeffrey Sachs. CXI is going deep into the fabric of modern economies to look at how AI will "disrupt" the lives of working people, to use the euphemism for the destructive innovation ideology of digitalization. Its analysis begins with the problems of GDP economies, meaning environments where bottom lines and growth are the only meaningful parameters. CXI takes issue with the term “artificial intelligence” itself, preferring the term “extended intelligence.”

> *John C. Havens, executive director of the Council on Extended Intelligence:* ***“A lot of the messaging right now is about new AI technology that defeats humans in one more area. It is not only disempowering, it is like, what are you all trying to do? Our council member Joichi Ito, head of the MIT Media Lab, has called what we are aiming to do ‘reducing reductionism’. He said instead of thinking about machine intelligence in terms of humans versus machines, we should consider the system that integrates humans and machines, not artificial intelligence, but extended intelligence.”***

Tackling just one aspect of the debate is a non-profit group of academics, technologists, and scientists based in the San Francisco Bay Area. AI4ALL is working on creating a more diverse base of scientists and workers in the field of AI. Its members recognize the potential dangers posed by the bias inherent to a system created by mostly white male engineers in the industrialized countries of the West (women currently make up only 12 percent of engineers working in AI[8]). As Fei-Fei Li, a chief scientist at Google, has said, “We all have a responsibility to make sure everyone—including companies, governments, and researchers—develops AI with diversity in mind. Technology could benefit or hurt people, so the usage of tech is the responsibility of humanity as a whole, not just the discoverer. I am a person before I’m an AI technologist.”[9]

The app industry is a prime example of the bias problem, as it churns out more and more solutions for problems affecting mostly affluent inhabitants of metropolitan areas, such as the desire to optimize the process of ordering a taxi or a pizza. The problem is not so much the flood of lifestyle products, but the impact on research and development in other areas. This is a similar to the pharmaceutical industry, in which the success of lifestyle drugs (think mood or virility enhancers) have redirected resources from potentially unprofitable areas such as research into eradicating malaria.

Supported by prestigious institutions including Stanford, Princeton, Berkeley, and Carnegie Mellon, AI4ALL sets up educational programs, mentorships, and creates pipelines for the placement of a diverse workforce in AI.

> *Tess Posner, CEO of AI4ALL:* ***“At this inflection point in AI, access is crucial. Right now, we have the opportunity to tackle big, long-term issues in this game-changing technology such as unconscious social bias. Increasing access to AI is important not only for altruistic reasons. Research out of Intel has proven that a more diverse US workforce would annually add $500 billion to the economy. With a diverse workforce, you see that people bring different experiences and thus different ideas to the table, increasing creativity and widening the spectrum of problems that get addressed in research and development. In fact, research predicts that increasing diversity and access to the innovation economy would quadruple the rate of innovation in the US.”***

It is the work of women, men, and organizations such as Max Tegmark and the Future of Life Institute, John Havens and the Council on Extended Intelligence, and Tess Posner and AI4ALL that is laying the groundwork of embedding the right values in the DNA of AI. The fact that these debates and efforts are coming now is crucial. To date there have been no "wake-up catastrophes," as nuclear physicist and philosopher Carl Friedrich von Weizsäcker once termed the human tendency to address problems only after a terrible event. It took the attacks on Hiroshima and Nagasaki, for example, for the public to understand the downsides of nuclear technology. And the human collective can be dangerously slow to respond to threats, as the climate change debate demonstrates. The good news is that the world seems to have realized relatively early that the power AI might unleash requires a thorough review, as well as the creation and institution of ethical guidelines, to steer its development.

PART 4: BEYOND THE DEBATE

The debate about establishing common values for the development of AI—and the efforts to do so—have three basic elements that give reason to be optimistic in principle.

- All debates and actions are geared towards a long-term future with few known outcomes, so all efforts are driven by vision, not planning.
- The creation of values is based on ethics, not morals, fortifying common guidelines against the interference of politics and ideology.
- Globally, even in the highly competitive environment of the US West Coast, there is a sense of common goals and the need for cooperation overriding business interests.

Despite these hopeful signs, the debates will remain heated and urgent. They have to be. The one single point that everyone—from the singularity enthusiasts to the doomsday prophets—agrees on is the fact that AI will be one of the most powerful technologies humankind has ever created. But since its progress has no clear trajectory right now, both hopes and fears are high. If extreme viewpoints continue to dominate headlines, it will remain hard to sift the realists from the extremists.

There are three schools of realist thinking right now. First, there is the technological realist, personified by John Cohn, computer engineer and "chief agitator" of the IBM Internet of Things division which uses the powerful Watson AI:

> *"Everybody in AI is starting to use the same sort of frameworks. That gives you the possibility of building in safeguards. You can't prevent some bad agent from circumventing those. But if you think about it, the fact that almost all the world's main practical AI is evolving as a kind of combined open-source thing is really phenomenal. I've never seen anything like this in the world. Because everybody sees the common good in those safeguards. The notion that when something like in the paperclip analogy runs away, you can build something like run-away watchdogs into the infrastructures that are used by most of the world. I am definitely cautiously optimistic about the rise of AI."*

Second, there are the societal realists, such as John Markoff, technology reporter of the New York Times, who grew up in Palo Alto, one of the central communities of Silicon Valley:

> *"It wasn't an accident that personal computing happened first in Silicon Valley. That was because of this cultural collision between counterculture and the micro-processor. Now it's interesting, the dip flipped some time in 2015. The bright shiny thing in Silicon Valley, which had been the social networks, became machine intelligence. It's just that the fire hose of venture funding went from here to here. The economy will continue to evolve, but what I'm quarreling with is the inevitability of crisis. We're going to reach a crisis point. There was a profound change in the economy with industrialization. I can see a similar transition. But I'm actually optimistic about this generation of scientists thinking about the consequences that work in America."*

And third there is the philosophical realist, most famously identified with Steven Pinker, bestselling author and cognitive scientist at Harvard University:

> *"Recent baby steps toward more intelligent machines have led to a revival of the recurring anxiety that our knowledge will doom us. My own view is that current fears of computers running amok are a waste of emotional energy—that the scenario is closer to the Y2K bug than the Manhattan Project. It's bizarre to think that roboticists will not build in safeguards against harm as they proceed. They would not even need any ponderous 'rules of robotics' or some newfangled moral philosophy to do this, just the same common sense that went into the design of food processors, table saws, space heaters, and automobiles. Would an artificially intelligent system deliberately disable these safeguards? Why would it want to? AI dystopias project a parochial alpha-male psychology onto the concept of intelligence. They assume that superhumanly intelligent robots would develop goals like deposing their masters or*

> ***taking over the world. But being smart is not the same as wanting something. Once we put aside the sci-fi disaster plots, the possibility of advanced AI is exhilarating—not just for the practical benefits, like the fantastic gains in safety, leisure, and environment-friendliness of self-driving cars, but for the philosophical possibilities. The computational theory of mind has never explained the existence of consciousness in the sense of first-person subjectivity. Imagine an intelligent robot programmed to monitor its own systems and pose scientific questions. If, unprompted, it asked about why it itself had subjective experiences, I'd take the idea seriously."***

Organizations are expected to spend $52.2 billion every year on AI-related products by 2021; PwC estimates AI could contribute $15.7 trillion to the global economy by 2030[10]. The next few years could amount to an AI "gold rush." In that scenario, the two main competing economic powers—the US and China—appear to have distinct advantages. Venture capital in the US takes more risks and pulls from a greater pool than European capital. And a tech culture that generally accepts fast failures allows companies and developers in the US to be more aggressively innovative. China, on the other hand, has few ethical or legal boundaries to constrain innovation, and the vast resources of a supportive government to draw upon.

While these advantages might seem like threats to other countries, they will only guarantee short-term success. As the current techlash shows, missteps and mistakes can significantly slow or halt progress. Companies and developers in Europe have a significant longer-term advantage right now: their will to build technologies based on clearly articulated values. As a consequence, they will become powerful forces in the continuing process of digitalization fueled by AI.

Germany in particular has some advantages, making it a robust competitor of the US and China, despite its lack of visionaries and grand projects.

> *Ulli Waltinger, founder and co-head of the Siemens Artificial Intelligence Lab:* ***"In Germany we have the great advantage of our industrial expertise and manufacturing know-how. The US uses AI and digitalization largely for service-oriented platforms like search engines and social media. In Germany, industry has started to implement, pilot, and deploy AI for various purposes, like prescriptive and predictive maintenance or the generation of digital twins, meaning the simulation of real-life processes, which can be then used for the optimization of production or processes. But there is another aspect of the German industry that will become an important factor. Security, trust, and reliability have always been hallmarks of the German market and its industrial ecosystem. Those values will become even more important in the entire lifecycle of AI-related products and processes."***

In summary, the lessons learned from the past decades of scientific and technological progress promise to shape the future of AI. There is a clear thrust of history making this the likely scenario. Just as the catastrophe of nuclear arms led to a cautious approach to bioengineering, the digitalization techlash and the change in attitudes and approaches it inspires will most likely lead to the reasonable development of AI.

> *Sarah Spiekermann, Chair of the Institute for Management Information Systems at the Vienna University of Economics and Business: "**Currently, engineers are given few guidelines on how to deal with the non-functional requirements of the systems they are working on, i.e., what values beyond dependability and security to build for. If there are ethical issues at stake, such as privacy, this is mostly delegated to legal departments. Having engineers work in an ethical manner is of course both time- and cost-intensive. Beyond facing these daily work challenges, technical innovation teams are now at a stage where they need to ask even bigger questions such as 'why do we want AI? Do we want to mimic human traits with technology? To reach truly ethical guidelines, we have to be aware that our vision of human nature has been questionably dominated by a school of negative viewpoints. It started with philosophers like Machiavelli and later Thomas Hobbes who in the 18th century stated, 'Man is wolf to man'. Such views on human nature continued with the idea of the immanent fallibility of humans in the philosophies of John Locke and Jean-Jacques Rousseau, who saw the negative side of humans as a justification for a strong state built on law, order, and education. This conception of human nature is not only embedded in our society; it is amplified by an approach to technologies like AI that sees humans as the weakest link in progress. Well-known developers publicly claim such things as humans being 'the last bug' in the system. In contrast, any value system for AI should be based on a healthy and trustful idea of humans. Only then will value-based engineering be possible.**"*

For Europe and Germany, the challenge will be to find and fortify their roles on the vanguard of this trajectory. If European and German companies set themselves up as leaders in defining the values that matter in the current debate, they will join the ranks of companies from the US and China in not just shaping, but dominating the future of AI technology. This will require both a clear vision of the values they hold, and the self-confidence and will to take on a leadership role. The debate about AI is now at the right stage to ask questions. The right answers will have to follow soon. Cautious optimism — by scientists and developers, by lawmakers, and by the general public—will remain the right mindset for proceeding into the future.

NOTES

[1] Hewes, A. (2012). SH Interviews Vernor Vinge.
Available at: https://bit.ly/2BRbPoh

[2] Vinge, V. (1983). First Word, p.10.

[3] Bolstrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence.

[4] Powell, A. (2018). Onward and upward, robots. The Harvard Gazette.

[5] EU GDPR (2018). Available at: https://www.eugdpr.org/

[6] Berg, P. (2008). Meetings that changed the world: Asilomar 1975. Nature.
Available at: https://www.nature.com/articles/455290a

[7] Future of Life (2017).
Available at: https://futureoflife.org/ai-principles/

[8] Adams, S. (2018). The Tech Unicorn That Went For Women Engineers: Here's How It Worked Out. Forbes.

[9] Yao, M. (2017). Meet These Incredible Women Advancing A.I. Research.
https://www.forbes.com/sites/mariyayao/2017/05/18/meet-20-incredible-women-advancing-a-i-research/#408088d926f9

[10] PwC (2018). PwC's Global Artificial Intelligence Study: Exploiting the AI Revolution.
Available at: https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html