Source: © Yuya Shino, Reuters

The Digital Future

Rules for Robots

Why We Need a Digital Magna Carta for the Age of Intelligent Machines

Olaf Groth / Mark Nitzberg / Mark Esposito

We stand at a turning point in human history, on the threshold of an unknown digital future. A powerful new technology, artificial intelligence (AI), permeates every area of our lives, largely thanks to advances in neural networks, modelled loosely on the human brain. Our societies and economies have become increasingly dependent on the use of artificial intelligence. A new set of rules is needed in order to ensure that freedom, inclusion and growth are safeguarded in the future. In other words, we need a digital Magna Carta for the age of cognitive machines.

Dawn of the Cognitive Age

Artificial intelligence can detect patterns in massive unstructured data sets.[1] In view of the increasing availability of data, it can improve the performance of companies, identify objects quickly and accurately, and enable ever faster decision-making, whilst minimising the disruptive influences of complex political and human circumstances. This constellation raises fundamental questions about the degree of human freedom of choice and inclusion, the significance of which will increase in the coming decades. There are, moreover, crucial differences in the attitudes and approaches of leading nations with regard to these issues. The current differences in the international value structure will intensify and the potential for social and geopolitical conflicts is rife.

In future, to what extent will humans – including the elites and representatives of all positions of power and levels of income – still be involved in decision-making processes? How can we govern this brave new world of 'machine meritocracy'?

In order to find an answer to these questions, we need to travel back 800 years. Upon his return from France, in January 1215, King John of England faced angry barons who wished to end his unpopular rule of *vis et voluntas* ("force and will") over the realm. In an effort to appease them, the King and the Archbishop of Canterbury brought 25 rebellious barons together to negotiate a "Charter of Liberties" that would enshrine a body of rights for the aristocrats to serve as a check on the King's discretionary power. After lengthy negotiations, an agreement was finally reached in June that provided greater transparency in royal decision-making, a louder voice for the aristocrats, limits on taxes and feudal payments, and even some rights for serfs. This was the famous Magna Carta. It, of course, remained an imperfect document, teeming with special-interest provisions of certain social classes. Yet, today we tend to regard the Magna Carta as a watershed moment in humanity's advancement toward an equitable relationship between power and those subject to it. Ultimately, it set the stage for the Enlightenment, the Renaissance and today's constitutional democracy.

Similarly, it is the balance between the ever-increasing power of the new potentate – the intelligent machine – and that of mankind that is at stake today. This is a world in which machines are increasingly involved in the creation of value, produce more and more everyday products, and in which human control over design and numerous other important aspects is being continuously reduced. As a result, our current work-life patterns will, in the long term, irreparably change. This technology, which we have ourselves created, will soon overtake certain cognitive abilities of humans, and thus increase their lead ahead of us in terms of productivity and efficiency at a breathtaking pace.

The consequences of this will become apparent in the following one to two decades. After all, it takes computers mere weeks, or often hours, to recognise complicated patterns in dozens or hundreds of data streams that have been generated as a result of centuries of scientific work and economic activity. And they do it with a precision and tirelessness that is far superior to anything that humans can offer. Acquiring knowledge and insight from this, and communicating decisions, is at the core of cognition, i.e. of thinking.

The cognitive ability of AI will transform human existence over the next 10 to 20 years.

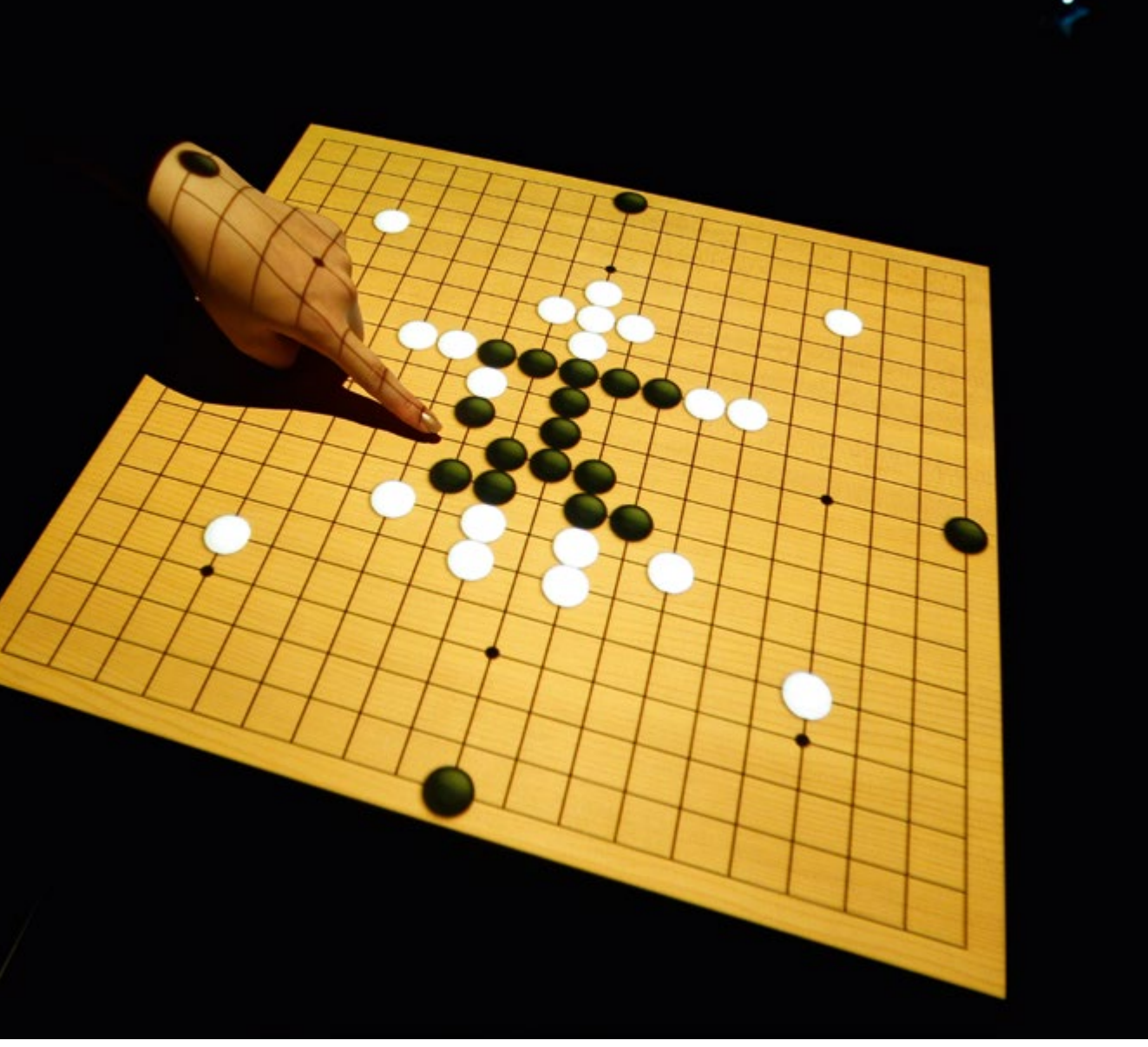
There is no doubt that machines are still decades away from replicating the human brain's intuitive ability to project – a capacity that has evolved over millions of years. With less data, we are still superior to machines, since it is in our DNA to be able to think outside the box, to be inspired, and to come up with ideas as a result of thoughts colliding. Then there is the human being's emotional intelligence, empathy, consciousness, moral understanding and ability to be intuitive, to interpret and to sense things. However, we should not fall into the trap of thinking that intelligent machines can never develop similar skills, which may not correspond exactly to our own, but which could circumvent or replace them to a certain extent. This would not be developed with a view to totally replace humans, but instead to enrich our lives. However, it also exposes us to dangers. The potential for disruption has been ignored mainly due to the fact that Hollywood has given AI so many attractive faces and voices. In the film "Her", AI in the form of Scarlett Johansson certainly stimulates our imaginations, but most realistic, critically-thinking people believe such a phenomenon is still a long way off. However, this critical reflection and realism has the drawback that it may lead to simple mockery of the hype coming out of the United States and China, rather than spurring people to take it seriously and get involved in shaping it.

From Homo Economicus to Homo Digitalis

Cognitive computers and intelligent machines are no longer the stuff of science fiction. There are already many good examples: for instance, demographic change and an ageing population have led Japan to the brink of a crisis, due to a shortage of nurses and care workers. Faced with a need for a million nurses, social workers, and so forth, the country has invested heavily in robot technology to make life easier for older people. Robots are not only used for heavy work such as lifting patients or shopping for the elderly, but also for social services such as simulated contact with pets and fellow human beings, which is offered by more or less realistic, animal-like and humanoid robots (see also the article on Japan in this issue). Meanwhile, the United Arab Emirates has recently set up a separate ministry for artificial intelligence and has started pilot projects using simple robocops to provide basic observation and information services. In Nigeria, applications such as Touchabl.com are being developed so that, for instance, even illiterate people can participate in social production processes, play an active role as consumers, and thus contribute to the economy and develop a digital voice.

However, matters becomes more problematic when personality traits, demographic profiles and a large part of people's social interactions are digitally evaluated and rendered public, thus leading to the violation of personal rights and privacy, as these concepts are understood in the West.

In the US, the UK, in China and in Russia, AI technologies such as facial recognition, speech processing and mood analysis algorithms are being used to prevent crime and terrorism. In this way, police departments in New York City and Los Angeles can track the crime risk of certain individuals and entire neighbourhoods, and deploy officers at the right time. The extent to which these AI applications are, in their entirety,



Go robot, go! Compared to chess, the board game Go poses a disproportionately greater challenge to AI. However, by now, computer programmes have surpassed humans even here. Source: © Kim Kyung-Hoon, Reuters.

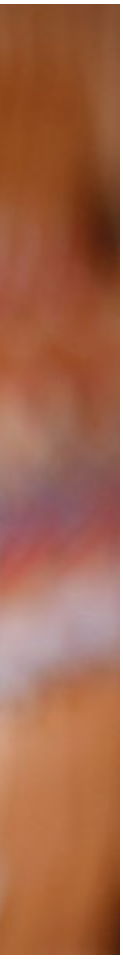
subject to privacy and data protection considerations varies considerably from country to country. In some districts of Los Angeles, for example, street cameras capture the faces of people who were in the area when crimes were committed. This, however, means that people who were not actually involved, but who have collected points through correlation, are also added to databases. The admissibility of this procedure is to be determined by a lawsuit being brought by the American Civil Liberties Union. There is a similar approach in New York City, where the New York Police Department (NYPD) uses algorithms supplied by technology companies to ascertain where criminal offences are most likely to occur in the city, so that they can then deploy police officers as a preventive measure. Courts in Wisconsin and Florida are also using predictive analytics to make judgments about what kind of risk a defendant poses, and accordingly, what bail should be set. Here, as in New York, the judges or their "intelligent" technologies cannot say exactly what logic was used to make a certain judgment, because the

algorithms are composed as neural networks that propagate and re-propagate conclusions between different levels of the network at high speed, but without being able to understand them. This goes too far, even for Americans who invoke the transparency of their legal system, so the Pentagon's Defense Advanced Research Projects Agency (DARPA) has launched an "Explainable AI" project.

In China, multi-billion AI systems are currently being developed, which provide every citizen with a public "trust rating".[2] China's internet behemoths and the Chinese government are developing AI-driven systems such as the Zhima Credit (Sesame Credit) programme, which uses data such as praise or complaints from fellow citizens and government agencies to create a ranking list based on a scoring system of up to 800 points for individuals. This calculation also includes basic demographic data. So, for example, a 28-year-old pregnant woman will enjoy a better "rating" than an 18-year-old who purchases a motorbike. Someone who has 700 Sesame Credits is considered to be extremely respectable, whereas a score of 300 may lead to social repercussions, such as not being able to book an international flight. The official aim is to counteract corruption and untrustworthy patterns of behaviour, and to create greater reliability in economic and interpersonal transactions. Anyone who has ever been to China has been

From Alexa to Barbie : It is no longer a thing of the future that everyday objects are surveying us.
Source: © Mario Anzuoni, Reuters.

confronted by the problem of attempting to buy genuine products, such as branded goods. In addition, China has so far lacked a mechanism by which to rate consumer creditworthiness, as guaranteed by credit rating agencies in the US or by the Schufa system in Germany, which is necessary for a functioning economy. However, it is unclear how this system is protected against misuse or how the Chinese consumer can correct mistakes and misunderstandings. Experts also suspect that this will lead to the formation of new social classes, because people with good ratings could shy away from those with poor ratings. However, the most serious consequence of these systems is state control and re-education for the purposes of obedience to the State. In a world in which not only people's criminal activity, but also their demographic characteristics and their interpersonal "rough edges" are closely examined, quantitatively and statistically evaluated and then made public, one has to ask how far we are from the digital branding iron. Participation is currently voluntary, but it is destined to become compulsory by 2020. How will this affect the economic existence and social dynamics of 1.5 billion Chinese people? And will the resultant shifts lead to international repercussions? Will Chinese companies also introduce these systems in developing countries that receive aid from China, for example as part of the new One Belt One Road policy?



The fact is, predictive analytics, AI and interactive robotics already have to be regarded as fixture, being an essential tool for governments and businesses. And it is also a fact that public and political debate, particularly cross-border dialogue, is lagging far behind technological advances.

What Do Machines Know?

However, all this is not limited to the security sector. In every area of our lives, machines are starting to make decisions for us. They recognise our patterns of behaviour and thinking, and those of supposedly similar people across the world. We receive messages that shape our opinions, outlooks and actions based on tendencies we have shown (or which other similar people have shown) in past behaviour. Whilst driving our cars, car manufacturers and insurance companies collect information about our behavioural patterns to offer us ever-improving navigation aids and increasingly autonomous vehicle technology, that makes road traffic safer and more comfortable. We enjoy more and more sophisticated, customised entertainment and video games, the makers of which know our socioeconomic profiles, patterns of movement, and cognitive and visual preferences – knowledge they use to determine pricing sensitivity. The latest development in Mattel's Barbie doll serves as a good example of this. Now withdrawn from the market, it stored information about how children played, responded and spoke on Mattel's servers, so that other adaptive and targeted services could be offered. These might include allowing parents to monitor their children remotely, providing information on children's social behaviour and imagination, and certainly aiding in the development of new products. All this, however, rather smacks of Big Brother, especially as regards commercial interests marketing the data of an unprotected and unknowing minor. Customers were not (yet!) prepared to tolerate such an invasion of privacy. But this is already happening in many ways: today, every smartphone collects such data, distributes it for advertising purposes or uses it for the new, aforementioned improved services and products that are intended to enrich our lives in an ever more targeted way. Scientific Revenue, a start-up located south of San Francisco, enables the developers of computer and mobile games to determine and project the gaming behaviour, context and price sensitivity of players with a view to setting individual prices for a game or in-game purchase. Of course, this is not entirely new. In many places around the world, when negotiating prices for goods and services, the socio-economic impression we give our counterparts plays a role in how the prices get set. AI-enabled digital platforms codify, amplify and scale these processes – but without the desired transparency. Such a degree of individualisation is, on the one hand, enriching and pleasant. On the other hand, we should be aware that

we are already relying on machines to "know what is right for us." And indeed, the machine may get to know us even better than we know ourselves — at least from a strictly rational and empirical perspective. But the machine will not so readily account for cognitive dissonances between that which we purport to be and that which we actually are. Reliant on real data from our real actions, the machine constrains us to what we have been, rather than what we wish we were or what we hope to become.

Personal Freedom of Choice

Will the machine restrict our individual freedom of choice and development? Will it do away with life's serendipity? Will it plan our existence so comprehensively so that we only meet people similar to ourselves, and thus deprive us of encounters and friction that force us to evolve into different, perhaps better human beings? There is tremendous potential for improvement in AI. Some of our personal decisions should in fact be driven by more objective analyses: for instance, a rational synthesis of the carbon footprint for different modes of transport, our schedules and socio-emotional needs could lead to more sensible decisions on environmental policy. A look at the divorce rates in most industrialised countries could lead to the conclusion that it would not hurt to get a few objective, analytical pointers about how sensibly we select a partner and who really is right for whom. After all, our self-image and aspirations do not always coincide with our real behavioural patterns – a phenomenon psychologists call "cognitive dissonance". In addition, more effective curricula and cognitive-adaptive teaching should be developed for different groups of pupils and students with different learning profiles. American AI-experts are already working on systems that can help us avoid food shortages and famines by integrating changes in factors like weather, soil, infrastructure and markets into complex models to provide timely relief. The number of useful applications is almost infinite.

Polarisation or Social Balance?

However, artificial intelligence could also polarise societies by pushing us further into virtual bubbles of like-minded people, reinforcing our beliefs and values without giving us the chance to review, defend, or possibly revise these views through occasional confrontation with dissenting parties. Last but not least, AI could also be misused for digital social engineering, creating parallel micro-societies. For example, room or apartment letting agencies in certain districts might only rent accommodation to tenants with a particular socio-political, economic or psychometric profile, or only rent properties from such providers.

AI is already being used by businesses in selecting their empolyees more rigorously during job interviews.

Businesses could also use AI to help them select their employees much more specifically during job interviews. This is done using algorithms that evaluate the video streams of the various candidates according to the behavioural criteria that are important to the company. This is already happening today to some extent. Promising start-ups such as HireView and Koru in the US are making great strides and are already well-established in the industry, with a customer base that includes Unilever, Urban Outfitters and Vodafone. As a result, these companies are able to provide a more objective method for analysing interpersonal conversations, which are often beset with personal biases. This serves as a counterbalance to the CV, which has turned out to be relatively ineffective in predicting a potential employee's professional success in new situations. It is also sometimes easier to recruit candidates who appear less "glamorous" against the backdrop of their previous experience but whose situational behaviour is better suited to

the company. This could be particularly desirable for minorities, members of which are often overlooked. Ultimately, this will increase an employer's short-term success rate, but it also raises the question of whether an overly narrow method of analysis might not also lead to the excessive homogenisation of the workforce, which, in turn, could restrict companies' longer-term strategic options.

What About Values?

Machines judge us on our *expressed* behaviour and values, especially those implicit in our commercial transactions, because these deliver tangible, hard data. However, they overlook other deeply held values that we do not necessarily express in our actions at the time and for which there are no digitised data points yet. It is difficult for artificial intelligence to grasp newly formed beliefs or changes in our values outside the readily codifiable realm. As a result, it might, for example, make decisions about our safety that compromise the wellbeing of others based on historical data in ways we might find objectionable in the moment. We are complex beings who regularly make value and priority trade-offs within the context of the situation at hand, and sometimes those situations have little or no codified precedent for an AI to process. It is conceivable that an animal-lover's decision to make their self-driving vehicle swerve to evade an animal on the road – thus increasing their own risk of injury – could change when they have children of their own. Will the machine respect our rights to free will, to the evolution of our values, and the privilege of occasionally reinventing ourselves?

Discrimination and Bias

Similarly, a machine might discriminate against people of lesser health or social standing because its algorithms are based on pattern recognition and broad statistical averages. Uber has already faced an outcry over racial discrimination when its algorithms used postcodes to identify which neighbourhoods its riders were most likely to originate from. What kind of people will AI most benefit? Will it favour the survival of the fittest, the most popular or the most productive? Will it make those decisions transparently? And what method of recourse will be available to us should we have to defend ourselves?

We cannot put the genie back in the bottle, nor should we try to.

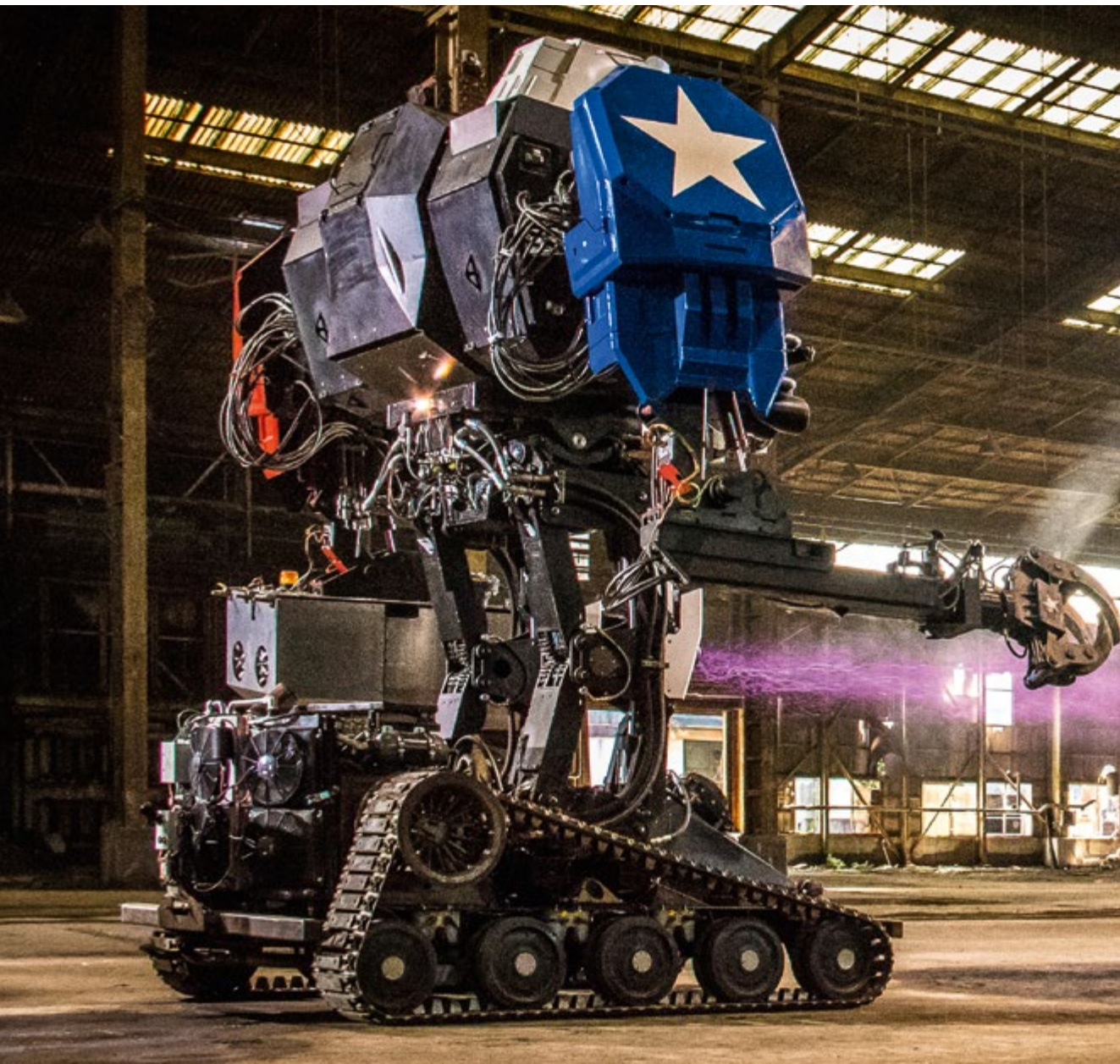
Moreover, a programmer's personal history, predispositions and unconscious biases – or the motivations and incentives provided by their employer – might unwillingly influence the design of algorithms and sourcing of data sets. Can we assume that AI will always maintain objectivity? What kind of AI systems are companies likely to develop? Will they act in the interests of their customers, partners, executives or shareholders? Will, for instance, a healthcare AI system jointly developed by technology firms, hospital chains and insurance companies act first and foremost in the patient's best interest, or will it prioritise financial return?

We cannot put the genie back in the bottle, nor should we try to – the benefits of AI will be transformative, possibly leading us to new frontiers of growth and development in human, social and economic spheres. One does not have to be a fan of utopian or dystopian science fiction to realise that we stand at the threshold of a fascinating and radical change in the evolution of humanity, unlike anything in the last millennium. Revolutions of this kind are rarely smooth. They are almost always chaotic, opaque, and fraught with ethical pitfalls.

A New Digital Ethics

The need for more ethics and responsibility in the digital realm was also clearly articulated during a three-day workshop on "The Future of Work", held at the University of California, Berkeley last November, and attended by representatives from the Organisation for Economic

Co-operation and Development (OECD), business and research. Some professional groups are already addressing this issue: for instance, the Institute for Electrical and Electronics Engineers (IEEE) has already drawn up a professional Code of Conduct. The Institute for the Future of Life, founded by MIT physicist Max Tegmark, is an association of leading scientists and entrepreneurs in the field of AI. Alongside a number of German and Austrian professors, its members include Elon Musk, Stephen Hawking, Ray Kurzweil, and Jaan Tallinn (founder of Skype). The Institute has drawn up the Asilomar Principles, named after a well-known conference centre in California. The Partnership on AI (PAI) is a technology industry consortium whose members include US internet giants Google and Microsoft, plus organisations such as Amnesty International and the American Civil Liberties Union. It is also drawing up guidelines, partly with a view to self-regulating as a way of anticipating future restrictive government legislation. The World Economic Forum (WEF), meanwhile, has just launched the first of a global series of new centres for the 4th Industrial Revolution in San Francisco, including an AI programme



to address these issues. Computer Science and Engineering faculties should also consider how to ensure their students and scientists take a responsible approach to AI design. In his interview with President Obama for WIRED[3] magazine, Joi Ito, director of the famous MIT Media Lab, pointed out that many of the most brilliant AI minds in laboratories such as his are not sufficiently attuned to the needs of people. This is partly because such "nerds" lack the patience to deal with human complexity, emotions and interpersonal politics, preferring to leave them out of the picture.

The AI industry is already working on its own guidelines in anticipation of possible restrictive government legislation.

What these existing initiatives lack, however, is – on the one hand – a truly global approach that addresses the complex mix of different values and definitions of ethics, and – on the other hand – the right blend of participants from different sectors of society, i.e. a multi-stakeholder approach.

A Magna Carta for the Digital Age

Cognitive technologies will not only determine the future of our economy, but also the future of our society. They influence what enters our minds, who knows our minds, with whom our minds cooperate, and how much thought our minds generate relative to machines. This has an impact on the whole fabric of society.

AI has to support mankind's growth as a whole as well as individual development so that people may realise their social and economic potential to the full. It is important to help people to deal with any uncertainties they may have regarding AI; they need to know that politicians will help them to prepare for the changes ahead. Furthermore, incentives must be provided for business and science to encourage them to implement both facets of AI in a purposeful and concerted manner. One way of doing this is through a digital Magna Carta for the AI-driven fourth industrial revolution. This should be a collectively developed charter of rights and values to guide our ongoing development of artificial intelligence. It should lay the groundwork for the future of human-machine coexistence and more inclusive human growth. Whether in an economic, social

Off to battle! Like any other technology, robots and AI can be used for the better or for the worse.
Source: © Michael Mauldin, MegaBots Inc., Reuters.

or political context, we as a society must start to identify rights, responsibilities and accountability guidelines so as to ensure inclusivity and fairness at the intersections of AI with our human lives.

Ideally, the Carta initiative should aim to constitutionalise a global multi-stakeholder institution for AI governance with a central team to track and analyse global developments in the area of AI (think-tank function) and discuss them in public plenary sessions (congress function). This should involve a culture of good faith collaboration amongst representatives from the private sector, governments and non-governmental organisations in the field of AI.[4]

Negotiations between the various economic, scientific, political and social interest groups ought to be conducted via a modern, open congress. This congress should allow for international, multi-sectoral participation, given that considerations relating to AI cross borders and overflow into every area of our lives and society. This requires that not only governments but also non-governmental organisations, academic institutions and business representatives come together at the same table to discuss the cross-border consequences of AI openly, rather than working at cross-purposes. In order to maintain incentives for all sides, the congress should aim to set rules that promote both innovation and equity.

This should be a new multilateral institution, which may act independently or under the auspices of the United Nations. It is vital that the institution's personnel and processes possess outstanding levels of digital proficiency in order to keep pace with the scientific and technical expertise of corporations, entrepreneurs and research laboratories. This is no mean feat, since AI talent is expensive in an environment where global internet companies pay salaries of several million dollars a year. Moreover, the congress should be inclusive, consisting of both physical and digital elements, so that the barriers and costs of participation remain low and dialogue can be driven forward rapidly. The mechanisms of the traditional institutions of the Bretton Woods system do not allow for this. They are equally unsuitable when it comes to ensuring a variety of key digital actors – such as China, India, Russia and Nigeria, all of whom are taking decisive steps towards shaping our digital future – have a formative voice, and are integrated in a constructive and critical manner. We should not be afraid of this openness, particularly in view of the unattractive alternative: just as in the case of the Asian Infrastructure Development Bank, which is managed from Beijing, we could end up with parallel institutions in which countries organise themselves into clubs. Such a situation would be in direct contrast to the global spread of AI technology by corporations and government agencies, as well as to the digital data flow itself, which crosses borders with ease.

With a view to reaching an agreement in the medium term, the focus of the Carta and congress should be, inter alia, the following issues:

1. What role should human freedom of choice play in the use of AI? How should the individual's freedom of choice and rights to privacy be protected? How will these protections be balanced against the needs of society?

2. How should we deal with actors who decide against the use of AI applications (e.g. granting an opt-out)?

3. To what extent can socio-political processes – such as elections, opinion-formation, education and upbringing – be supported by AI and how can the harmful uses of AI be prevented?

4. How can we effectively counter the corruption or falsification of data sets, as well as the potential discrimination against individuals or groups in data sets?

5. To what extent should policies and guidelines delineate the humane and nature-conform use and / or containment of AI?

6. How much importance should be placed upon social and societal benefits in the research, development, promotion and evaluation of AI projects?

7. How can the promotion and training of employees for new employment and personal growth opportunities be integrated into AI-driven automation of production and work processes?

8. How can effective, continuous exchange between different stakeholders be facilitated through AI?

9. What type of permanent international institution do we need that will provide the early foresight thinking, debate forum, and governance mechanisms needed to achieve responsible human and economic growth through AI?

It will not be easy, but unless we have a dialogue on these issues we will not establish sufficient trust in AI within global society so as to capitalise on the amazing opportunities it could afford us. It is only by agreeing upon a set of rules that we will be in a position to jointly steer the future of AI, and to ensure that the social compatibility of these revolutionary new technologies is internationally guaranteed and not only at the service of profit, power and geopolitical interests. Given the significant global capacities, such as the scientific and entrepreneurial talent in this field, if we did not dare to take this step, we would be missing a unique opportunity: to ensure a fair, value-based development of mankind as we step into the fourth industrial revolution – a revolution of cognition.

–translated from German–

Dr. Olaf Groth, Ph.D. is CEO of Cambrian.ai and Professor of Strategy, Innovation and Economics and Director of Digital Futures at Hult International Business School as well as Visiting Scholar at the UC Berkeley Roundtable on the International Economy.

Dr. Mark Nitzberg, Ph.D. is Executive Director of the Center for Human-Compatible AI at the University of California at Berkeley and serves as Principal and Chief Scientist at Cambrian.ai.

Dr. Mark Esposito, Ph.D. is co-founder of Nexus FrontierTech and Professor of Business and Economics at Hult International Business School.

1 Parts of this article have already appeared in the Harvard Business Manager Germany online blog (http://harvardbusinessmanager.de, Oct 2017).
2 Botsman, Rachel 2017: Big data meets Big Brother as China moves to rate its citizens, Wired, 21 Oct 2017, in: http://buff.ly/2l4rzMj [23 Feb 2018].
3 Dadich, Scott 2016: Barack Obama, Neural Nets, Self-Driving Cars, and the Future of the World, Wired, 11/2016, in: http://buff.ly/2dC2AXY [23 Feb 2018].
4 The authors have started to work on a concept for a "Cambrian Congress", which will facilitate both the potential Cambrian-like explosion of opportunity and mitigate the accompanying risks.