Artificial Intelligence

Ethics, governance and policy challenges

Report of a CEPS Task Force

Andrea Renda

Artificial Intelligence

Ethics, governance and policy challenges

Report of a CEPS Task Force

Andrea Renda

Centre for European Policy Studies (CEPS)
Brussels
February 2019

The Centre for European Policy Studies (CEPS) is an independent policy research institute based in Brussels. Its mission is to produce sound analytical research leading to constructive solutions to the challenges facing Europe today. This report is based on discussions in the CEPS Task Force on "Artificial Intelligence: Ethics, Governance and Policy Challenges". The Task Force, chaired by Andrea Renda, was composed of authoritative scholars, industry experts, entrepreneurs, practitioners and representatives of EU and international institutions. The group met on four occasions in the course of 2018. The views expressed do not necessarily represent the opinions of all the Task Force members, nor were they presented by any of the participants (unless explicitly mentioned in this report). A list of members and guest speakers appears in the Annex. The views expressed in this report are those of the author writing in a personal capacity and do not necessarily reflect those of CEPS or any other institution with which the members are associated.

ISBN 978-94-6138-716-5

CEPS
Place du Congrès 1, B-1000 Brussels
Tel: 32 (0) 2 229.39.11
e-mail: info@ceps.eu
internet: www.ceps.eu

Contents

List of Tables, Boxes and Figures

Glossary of Abbreviations

aaS	as a Service
AI HLEG	High Level Expert Group on Artificial Intelligence
AI	Artificial Intelligence
API	Application Programming Interface
API	Application Programming Interface
ARPA	Advanced Research Projects Agency
B2B	Business to Business
B2C	Business to Consumer
CLAIRE	Confederation of Labs of Artificial Intelligence Research
CPU	Central Processing Unit
DG EAC	Directorate-General for Education, Youth, Sport and Culture
EGE	European Group on Ethics in Science and New Technologies
ELLIS	European Lab for Learning and Intelligent Systems
ESIR Group	Economic and Societal Impact of Research and Innovation
ETSI	European Technical Standards Institute
EuroHPC	European High-Performance Computing Initiative
G2C	Government to Citizens
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GPT	General Purpose Technology
GPU	Graphics Processing Unit
HPC	High-Performance Computing
IEEE	Institute of Electrical and Electronics Engineers
IIC	Knowledge and Innovation Community
IoT	Internet of Things
ISP	Internet Service Provider
KPI	Key Performance Indicator
LAWs	Lethal Autonomous Weapons
OEM	Original Equipment Manufacturer
PEAS	Performance, Environment, Actuators and Sensors
P2P	Peer-to-Peer
PSI	Public Sector Information
RegTech	Regulatory Technology
SupTec	Supervisory Technology
TFP	Total Factor Productivity
TPU	Tensor Processing Unit
ZKP	Zero-Knowledge Proofs

INTRODUCTION: THE PROMISE OF ARTIFICIAL INTELLIGENCE

Like an unannounced guest, artificial intelligence (AI) has suddenly emerged from nerdy discussions in university labs and begun to infiltrate larger venues and policy circles around the globe. Everywhere, and particularly in Europe, the debate has been tainted by much noise and fear, as evidenced in the European Parliament's resounding report on civil law rules for robotics, in which Mary Shelley's Frankenstein is evoked on the opening page (European Parliament, 2016). At countless seminars, workshops and conferences, self-proclaimed "experts" voice concerns about robots taking our jobs, disrupting our social interactions, manipulating public opinion and political elections, and ultimately taking over the world by dismissing human beings, once and for all, as redundant and inefficient legacies of the past.

Part of this discussion comes with an underlying mantra: "AI is different": it is not like the Internet, not like electricity, not like the industrial revolution, not like oil and not like the invention of the wheel. Accordingly, so the gospel goes, we need new laws, new rules of conduct, new criteria for interacting with machines and a lifeline in case they decide to take over. Through the AI looking glass, the world suddenly seems a more dangerous place, and the Eldorado of today's society melts under the heat wave of smart autonomous robots. The neo-Luddites merge with those who are simply fearful, and the glass inevitably appears half empty.

Is this view justified? A first look suggests that the promise and challenges of artificial intelligence are perhaps less disruptive and probably more boring than talking about singularity, cyborgs and "robo-cene". But a closer look shows enormous promise, as will be shown in this report. The promise of AI is easy to spot if one considers two fundamental starting points. First, as AI arrives on our planet, it finds a society that is making progress in terms of life expectancy and the eradication of poverty and famine, but one that is also fraught with contradictions and inequality, with unsustainable production and consumption patterns as well as deteriorating social relationships. And while digital technology has evolved over the last half century, most of these trends are largely independent of the Internet, let alone artificial intelligence.

The question then becomes: Will the growth of AI exacerbate the contradictions of modern society? Or, can AI help us build a better world?

Second, and in a related vein, our comprehension of what AI is and what it can do is still in its infancy, even if AI has already become pervasive in sectors such as digital platforms, banking, e-commerce, insurance, healthcare, energy, defence and cybersecurity. As we take our first steps in this blossoming new world, we can still decide how AI can help us promote a better society and a more sustainable future. In other words, we have the chance to approach policy choices in the best possible way: by asking the right questions, at the right time and in the right sequence. This is indeed what we attempt to do with this report, which synthesizes six months of work by the members of the CEPS Task Force on Artificial Intelligence. We have called upon (real) experts from academia, industry and policy-making, entrepreneurs and civil society representatives to present their views in the course of four highly interactive meetings. This allowed the Task Force to collect an extraordinarily rich blend of views on where AI can lead Europe and the world, and what we can do about it today.

This report, which attempts to consolidate in one single document all these ideas and discussions, is accordingly structured as described below.

Part I is dedicated to sharpening our definition of AI, understanding its level of development and possible future evolution, and placing it in context by defining the whole evolving stack of technologies and applications that are surrounding AI, such as high-performance computing, big data analytics and the Internet of Things. We present the main findings of the Task Force in terms of the need for responsible, trustworthy AI, as well as the imperative to link AI to the global debate on building a more sustainable future.

Part II focuses on the EU perspective and defines a vision for AI, rooted in the complementarity between man and machine, by asking "what can AI do for Europe?", rather than "what can Europe do for AI?" This gives us an opportunity to reflect on the tensions that AI is likely to bring into the legal system, as well as into the overall architecture of EU policy; and leads us to propose that AI develops in a way that fits the thrust and direction of Europe's 2030 Agenda. We then turn to the current debate on the Draft Ethics Guidelines on AI, presented in December 2018, and propose that the European Commission publishes guidance for AI developers, vendors and distributors, as well as organizations deploying AI, explaining and collecting good and bad practices in areas such as the selection, sampling, curation and cleaning of data; the design of algorithms; machine training and feedback; as well as the release of algorithmic outputs and responding to undesirable outcomes and impacts. This part also discusses whether policy changes will be needed in order to create more legal certainty in Europe as AI gradually becomes more pervasive.

Rather than advocating radical change, this section looks at existing legislation to spot cases in which rules may become obsolete as a result of the emergence of AI-enabled systems and environments. Finally, we discuss possible industrial and innovation policy scenarios that Europe could consider in order to boost its competitiveness in the AI domain. These include a discussion of a possible "CERN for AI" and the launch of a "Mission IT", which orchestrates various streams of education, research and innovation with a view to nesting AI in a more sustainable vision of Europe's future.

At the end of the report, in Part III, we summarize our main policy recommendations. A full list of Task Force members and guest speakers can be found in the Annex.

PART I. ARTIFICIAL INTELLIGENCE: DEFINITION, LIMITS AND OPPORTUNITIES

The expression "artificial intelligence" implies the use of man-made techniques (Latin meaning of *artificialis*) to replicate the ability to "read inside" (*intus legere*) reality. This initial definition is technology-neutral, but it may be of little use if not further qualified. As a matter of fact, existing applications of AI mostly focus on three main functions: optimization, search/recommendation and diagnosis/prediction. AI is helping humans to optimise logistics and supply chains, better diagnose diseases, predict and prevent epidemics and fully tailor recommendations to the peculiar tastes and needs of end users. More generally, together with other digital technologies, AI is helping humans build a "fifth element" after air, earth, water and fire[1]: a data layer that increasingly surrounds us, gradually virtualizing our environment and multiplying our possibilities as mankind. All current applications belong to the specific domain of "narrow" AI. **Nothing in current AI developments suggests that AI will move towards developing human-like perception and awareness, or sentience, thus leading towards so-called "artificial general intelligence" in the immediate future**. Our discussions with experts in this field confirmed this finding: hence, the words "artificial intelligence" should not be taken literally: in most cases, what we call AI has nothing to do with general human intelligence and awareness.

This observation, of course, does not lead to a *capitis diminutio*. Artificial intelligence can be extremely disruptive, empowering, challenging and unpredictable, as is elaborated below.

[1] The four elements, as proposed by Empedocles – earth, water, air and fire – frequently occur in classical thought; Aristotle added a fifth element: aether.

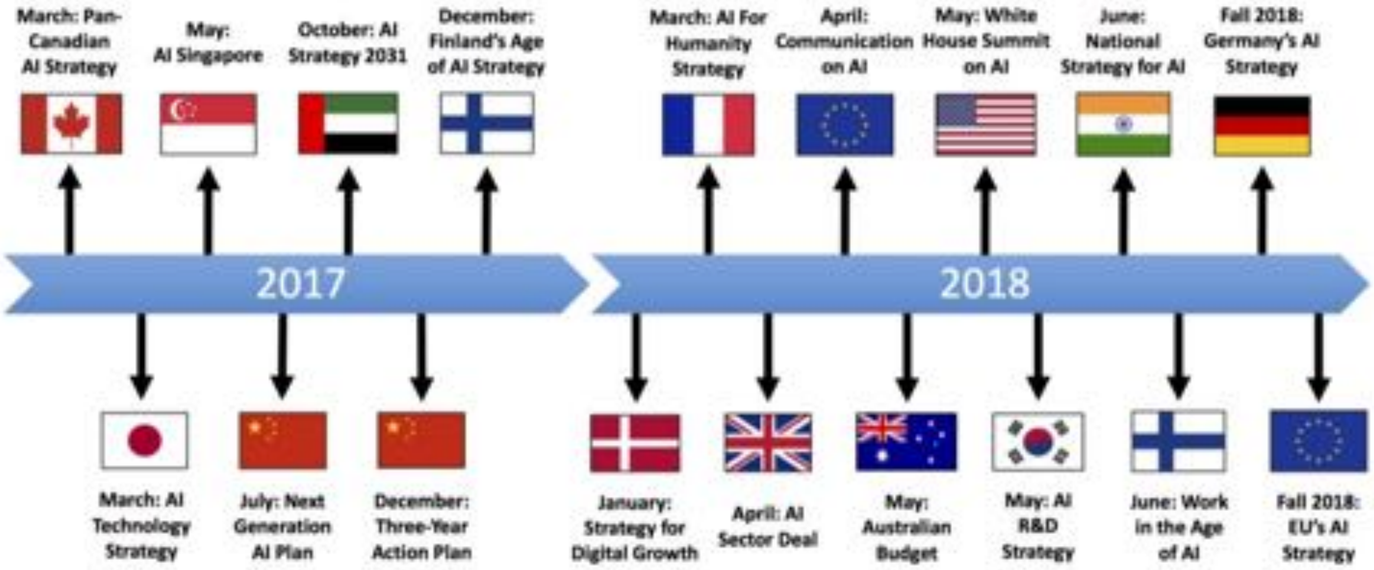
- ***AI can be disruptive*** since it is at the core of an emerging new stack, composed of big data and often also the Internet of Things (IoT), which is pervading many markets sweeping away or profoundly affecting incumbent business models. This development is leading humans towards new ways of working, interacting and building social relationships. AI's impact on productivity, as will be explained below, is still disputed, but is largely seen as more than simply a promise in the medium term: scholars are gradually abandoning gloomy prophecies of a replica of the "Solow Paradox" to embrace a vision in which AI becomes a game changer for total factor productivity and growth, by gradually rising as a third pillar of production, together with labour and capital.
- ***AI is empowering*** since it provides human beings with extremely powerful tools that, if properly used, can augment human intelligence and lead us onto more sustainable and desirable paths. AI can also gradually expand our "soft ethics" space, enabling more widespread education, better self-awareness, overall better health and life expectancy, and the possibility to navigate more sensibly through the "information envelope" that increasingly characterises society.
- Some ***AI developments can also pose challenges***, as in any powerful general-purpose, dual-use technology, since they can exacerbate existing societal biases and create new ones, deepen inequality, weaken competition and democracy, discriminate against minorities and entire social groups, and generally contribute to the ongoing deterioration of trust in modern societies. AI is also leading to a whole new generation of autonomous weapons and countless variants of extremely dangerous cyberattack tactics, including "deep fakes". This of course does not mean that AI is evil per se, but that humans could rely on AI to realise both virtuous and malicious goals, including building more deadly weapons and breaking security walls.
- Finally, ***AI can be unpredictable*** since existing algorithms can use unsupervised deep learning and neural networks in ways that generate outcomes that surprise even the original developers. This does not imply that AI is developing its own intelligence that departs from the goals and tools given to it by developers: however, these techniques instil an element of randomness and uncertainty in the way machines use data to reach optimising decisions.

Against this background, AI developers themselves, and increasingly also corporations and governments around the world, have been looking for ways to ensure that the positive disruption and empowerment effects of AI prevail over the potential negative effects. A global dialogue on AI has emerged, which revolves around countless ethical codes and declarations, from the "Asilomar principles" to the "Declaration of Toronto" and the "AI for Good" initiative;

corporate ethical principles developed by companies like Google, SAP, IBM, Microsoft, Deutsche Telekom, Telefonica referring to similar terms such as "responsible AI", "trusted AI", "trustworthy AI"; guidance for corporate practices developed e.g. by Accenture on tools such as algorithmic impact assessment, or by IBM with its AI Fairness 360 tool; government manifestos such as the Villani report, the Declaration of Montreal, the European Group on Ethics in Science and New Technologies (EGE) statement and the current draft ethical guidelines on Artificial Intelligence, the Chinese strategy on AI, the UAE strategy, the Indian strategy, etc. and full-fledged regulatory initiatives such as the EU GDPR (General Data Protection Regulation), the EU proposed Platform-to-Business regulation, etc. See Figure 1 for a timeline showing the adoption of national AI strategies by major countries across the globe in the period 2017-18.

Some of these documents aim at setting global principles or global standards governing AI; others at shaping corporate practices to enable compliance with established principles; and others at achieving industrial competitiveness, or sustainable development. Before we venture into this crowded space, it is important to take a step back and provide some basic information on, and definitions of, AI as we know it today.

Figure 1. National AI strategies


March: Pan-Canadian AI Strategy
May: AI Singapore
October: AI Strategy 2031
December: Finland's Age of AI Strategy
March: AI For Humanity Strategy
April: Communication on AI
May: White House Summit on AI
June: National Strategy for AI
Fall 2018: Germany's AI Strategy
2017
2018
March: AI Technology Strategy
July: Next Generation AI Plan
December: Three-Year Action Plan
January: Strategy for Digital Growth
April: AI Sector Deal
May: Australian Budget
May: AI R&D Strategy
June: Work in the Age of AI
Fall 2018: EU's AI Strategy

Source: T. Dutton (2018), "An Overview of National AI Strategies", Medium, 28 June (https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd).

1. A BASIC DEFINITION OF AI AND RELATED AI SYSTEMS

There are countless definitions of AI, and members of the CEPS Task Force have entertained a number of variants in the course of their meetings over the past year. For example, Mark Nitzberg AI Research Director at UC Berkeley, adopted a very simple definition: "narrow AI" gives the appearance of intelligent behaviour, while "general AI" matches human performance in all tasks. Czech Technical University Professor Michal Pěchouček defined AI as a family of technologies and scientific fields that allows/studies i) automation, ii) acceleration and iii) extreme scalability of human perception, decision making and reasoning. Niels J. Nilsson (2010) defined AI as "that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment". IBM's Francesca Rossi presented at CEPS an articulate, systemic position by defining AI as a family of techniques aimed at building machines that can gather input from the external environment, process it based on a given set of instructions and find ways to pursue the given goal through actuators. Finally, the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) created a sub-group tasked with reaching a definition of AI, chaired by Francesca Rossi. The group refers to AI as "systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."[2]

What this latter definition correctly captures is that **AI never appears in isolation, but as part of an IT system**, in which AI features as a "complementor" (Shapiro and Varian 1998). Sometimes this system is dedicated to a specific function, as in the case of robotics; in other cases, AI works as part of a more general-purpose system, as is the case of recommendation algorithms nested in services like Netflix, marketplaces like Amazon, Zalando or eBay, or search engines like Google. In any event, AI needs some form of hardware, in particular

[2] See "A definition of AI: Main capabilities and scientific disciplines", AI HLEG, 18 December 2018 (https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december.pdf).

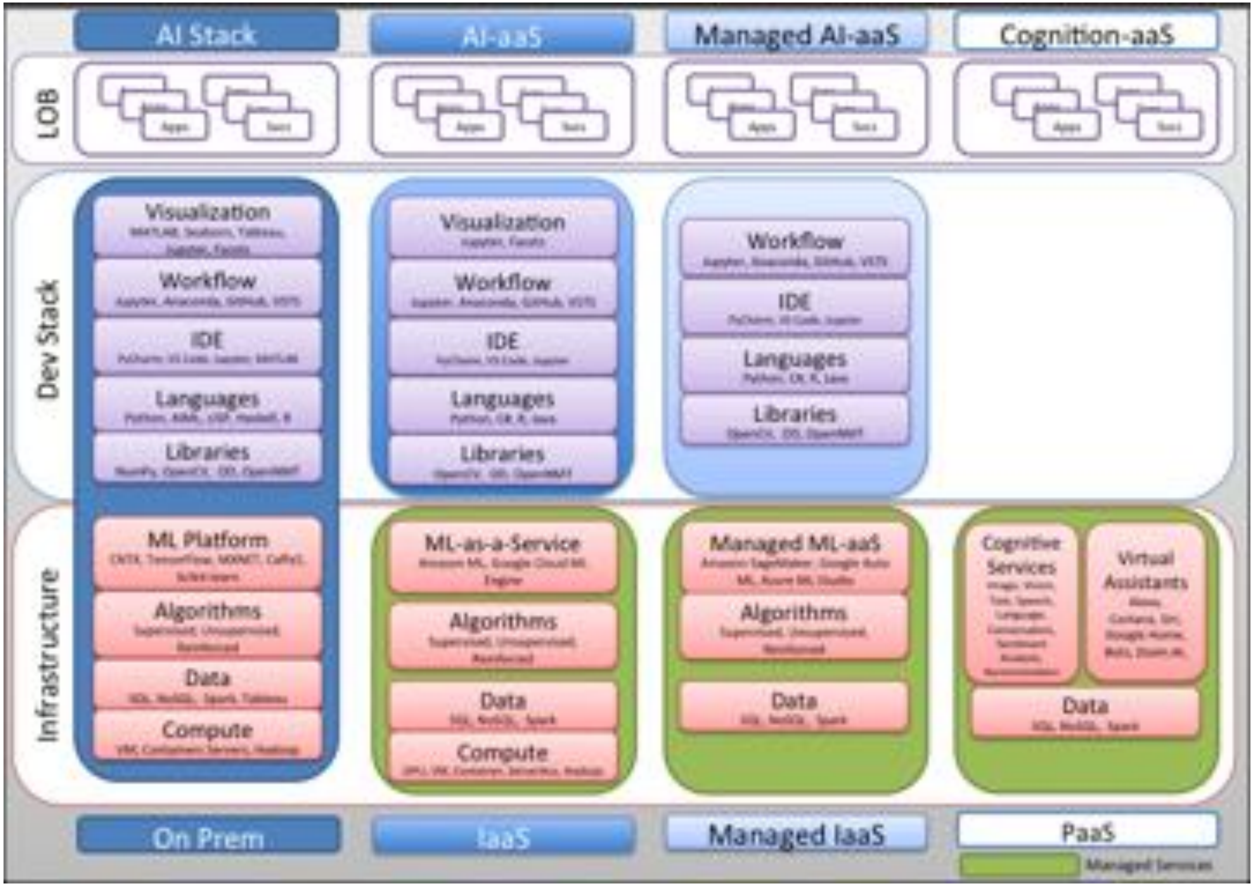
computing capacity, as well as data storage hardware and connectivity, for data acquisition (e.g. through sensors, or through direct typing of instructions, or communication with other machines) and for supporting actions through specific actuators, where appropriate. Moreover, AI needs software for the elaboration of data, their interpretation through data libraries, various types of algorithmic processing and decision-making, and for achieving interoperability with other systems, whether AI-enabled or not, and with humans.

These elements can be easily portrayed as **a layered "stack",** in which the hardware and connectivity components support the software, applications and service layers.[3] Figure 2 shows an example of "AI stack". As shown in the picture, the infrastructure requirement depends significantly on whether the AI-enabled system is used on premise, or in various "aaS" (as a Service) variants, which include different levels of combinations of hardware and software, as well as of what is used remotely and what is run locally. In many cases AI is used as an added functionality to existing cloud-based platforms, used "as a service" by customers through cloud-based access. AI is, thus, simply another feature at the application layer of the evolving layered stack used in all computer science. In other cases, AI uses dedicated hardware to perform its functions, and as such becomes the central "brain" of an ad-hoc, layered system that has dedicated hardware, such as specific computing power and specific "actuators", which enable the software to interact with its environment.

An intuitive example is that of the self-driving car, which requires a significant amount of hardware – including the mechanical components, the Lidar sensors, the cameras, sensors for cross-traffic alert, sensors for parking assistance – and a significant amount of software, including a general operating system, as well as dedicated software that enables ad-hoc functions such as parking assist, emergency braking, rear collision warning, blind spot detection, etc. In that environment, the AI system chiefly depends on its ability to collect data from the environment (through sensors and cameras), but also from available information infrastructure, e.g. vehicle-to-vehicle connectivity or V2V; vehicle to environment connectivity, or V2E; or vehicle to infrastructure connectivity, V2I). Data collected through existing maps, or directly from the environment including direct observation or V2X communication are elaborated by the system, and the car then decides on the best course of action given its underlying instructions and overall goals. The car then uses hardware to interact with the environment (so-called "actuators").

[3] Russell and Norwig (2009) observe that "AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."

Figure 2. Examples of AI stack



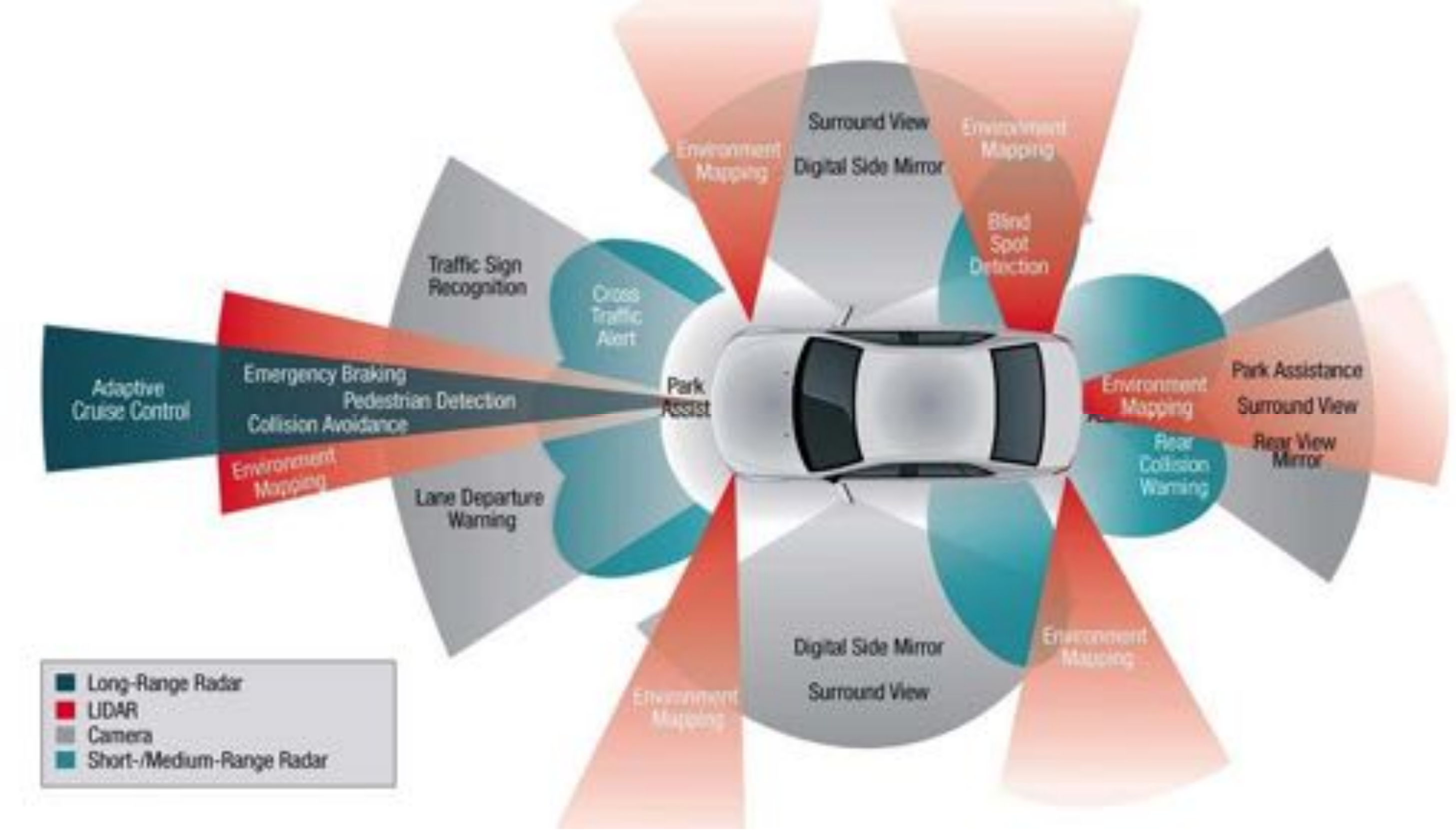
Source: Medium (https://medium.com).

Often these abilities are grouped in the term PEAS, which stands for Performance, Environment, Actuators and Sensors. Figure 3 offers a diagram of the complex set of radars and sensors that are enabling self-driving cars to appear on our roads today.

More generally, complex information systems are normally built as layers of so-called "complementors", which incorporate hardware, middleware and software components aimed at achieving both interaction with the external environment and elaboration of information. The typical properties of these system goods, as recognised in the academic literature, are: i) *modularity*, i.e. made of separate complementors that can be replaced as individual modules, provided that compatibility with the whole system, and in particular with the operating system that governs the device, is satisfied; ii) *interoperability*, i.e. the operating system that runs the whole system works with modules that are compatible with its application programming interfaces (APIs); and iii) *scalability*, indicating the capability of information systems to handle a growing amount of work or the potential to be enlarged to accommodate that growth.

That said, the representation of AI as a component in more complex information systems is useful for the purposes of our report, as well as for designing policy around AI. As a matter of fact, AI today is mostly used as an add-on to previous layered information systems, adding more "intelligence" to the application and service layers of the pre-existing stacks. Such intelligence is often very data-hungry, and as such requires powerful hardware, both for data collection and for data processing. In this respect, both the Internet of Things (IoT) and High-Performance Computing (HPC) are important components of AI-enabled systems. Without adequate hardware to sense the environment and interact with it, and without sufficient computing capacity, introducing sophisticated AI algorithms would be akin to driving a Ferrari to go to the supermarket or installing a smart grid in a "dumb" electricity network. Similarly, advances in nano-technologies and in the miniaturization of chips are essential to enable AI to be embedded in robots and other devices, including for example medical devices. It is very important to keep this in mind, since the promotion of AI in Europe through policy measures cannot ignore the relevance of the complementors that enable the full realisation of AI's potential.

Figure 3. Radars and sensors in a self-driving car



Source: Texas Instruments.

Box 1. Does an AI agent need to be rational?

Current AI systems mostly embed so-called "simple reflex agents", which select an action based on the current state only, ignoring historical data or past experience. Model-based reflex agents differ as they can act in partially observable environments by constantly updating their (static) representation of the world. A further evolution is represented by goal-based agents, used in cases where knowing the current state of the environment is not enough: the agents can combine the provided goal information with the environment model, to choose those actions that can achieve the given goal. Utility-based agents are an improvement over goal-based agents. These agents choose the action that maximizes the expected utility, after weighing both benefits and costs: they are thus very similar to the *homo oeconomicus* in economic theory. But the state of the art in AI goes beyond all these types of agents and implies the development of so-called "learning agents", which are based on the original definition given by Alan Turing. As agents become more complex, so does their internal structure, allowing for various forms of internal state representation.

Within the context of information systems, AI enables more complex decision-making, based on criteria that are, at least initially, provided to the machine by human beings. Russell and Norwig (2009) observe that AI "refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals". For example, rational decision-making, e.g., could be a goal for the AI developer, and as such the calculation of expected benefits and costs associated with a given action can be embedded in the system.

However, full rationality does not seem to be an essential element of an AI system: an AI developer could also try to replicate rational biases, such as impulsiveness, framing or hyperbolic discounting, to enable better interaction with human beings.[4] An AI system can also be taught to act in conditions of imperfect information, and as such can be trained to act with "rational ignorance" or in a more risk-averse way. Accordingly, the requirement of rationality does not seem to be needed in a definition of AI, even if ensuring that AI-enabled machines behave rationally will often be a clear goal of the developer.[5]

Also, rationality should not be intended as a process but rather as an outcome. There is no need for an AI system to replicate the same process followed by the human brain, based, i.a., on neurons and synapses. While neural networks are being used in deep learning processes, they are only one out of several possible ways to develop AI (see Silver et al. 2016). Postulating that AI should seek to mimic the functioning of the human brain would be also complicated if one considers that our brain functions are still relatively obscure for neuro-scientists (Adolphs 2015). At the same time, pretending that AI replicates or mimics the outcome of human decision-making would also imply that all the biases and imperfections of our decisions would be replicated in the AI system, and this, too, would be undesirable in many circumstances.

[4] Some scholars have proposed incorporating that quality into self-driving cars; see Renda (2018a).

[5] Within artificial intelligence, a rational agent is typically one that maximises its expected utility, given its current knowledge.

1.1 What AI can do: Main techniques and use cases

AI is already being massively used in a number of areas and **can be broken down in many sub-domains and techniques**. These include search and planning; knowledge representation and reasoning;[6] machine learning, which has led to AI breakthroughs in fields such as search and product recommendation engines, speech recognition, fraud detection, image understanding, etc.; multi-agent systems; robotics; machine perception, including computer vision and natural language processing; and more.[7]

In particular, *machine learning* accounts for approximately 60% of current investment in AI-related R&D: it extracts patterns from unlabelled data (unsupervised learning), or efficiently categorizes data according to pre-existing definitions embodied in a labelled data set (supervised learning).[8] Developers feed machine-learning systems large amounts of data, then the system finds the hidden relationships and uses reinforcement to improve its performance automatically. Machine learning is used in Google's search algorithm, digital advertising and online personalization tools (e.g. the Amazon and Netflix recommendation engines; or the Facebook newsfeed). Machine learning also extends into quantitative processes such as supply-chain operations, financial analysis, product pricing, and procurement-bid predictions. Today, nearly every industry is exploring or utilizing machine-learning applications.

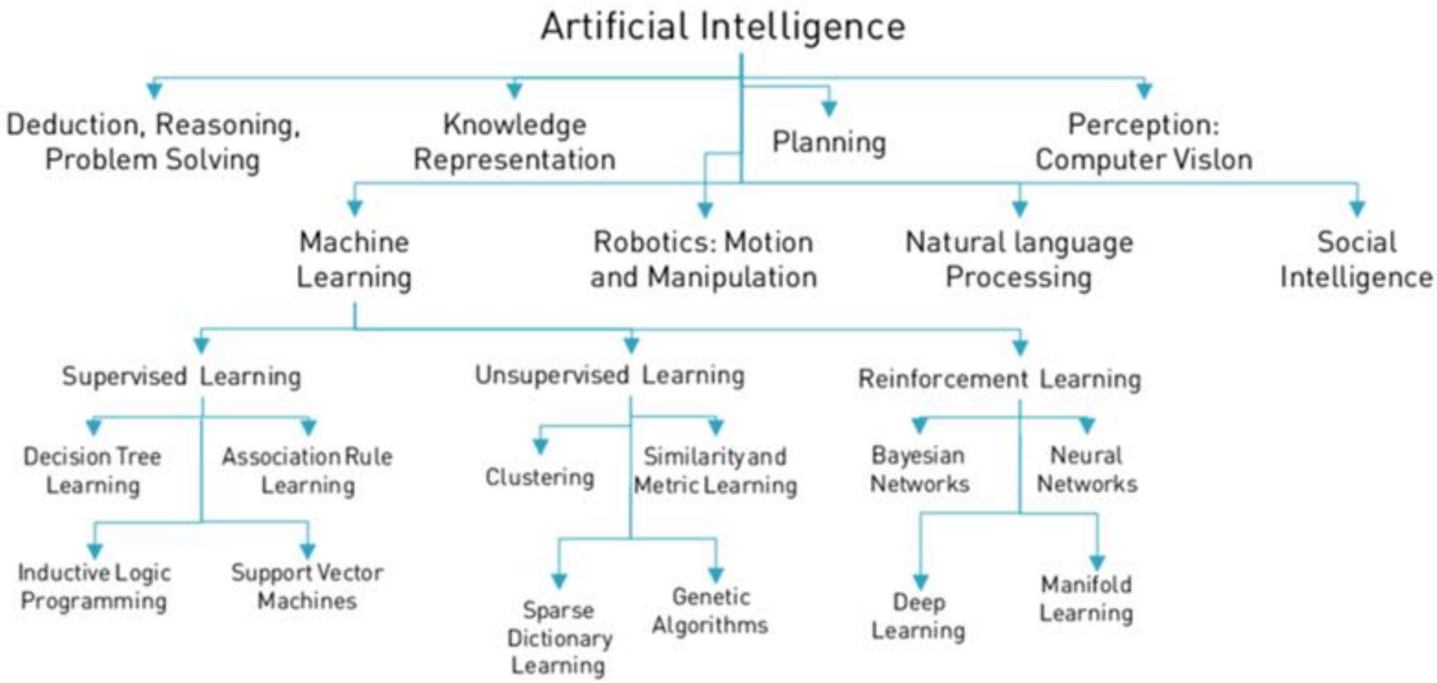
Within this domain, *deep learning* uses additional, hierarchical layers of processing (loosely analogous to neuron structures in the brain) and large data sets to model high-level abstractions and recognize patterns in extremely complex data. Deep learning has made speech understanding a reality on our phones and in our kitchens, and its algorithms can be applied to a wide array of applications that rely on pattern recognition. These tools are made available today by large corporations (Google's TensorFlow, Microsoft's Control Toolkit, and many other AI tools are indeed released on a free and open-source basis) and operate on common computer hardware. See Figure 4 on the following page for a classification of AI approaches and domains.

[6] The IBM Watson program, which beat human contenders to win the Jeopardy challenge in 2011, was largely based on an efficient scheme for organising, indexing and retrieving large amounts of information gathered from various sources. See https://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/

[7] Some of the most sophisticated AI systems use a combination of these techniques. For example, the AlphaGo programme that defeated the human champion at the game of Go used multiple machine learning algorithms for training itself, and also used a sophisticated search procedure while playing the game.

[8] Presentation to the CEPS Task Force by Michal Pěchouček (Czech Technology University).

Figure 4. Classification of AI approaches and domains



Source: Ajit Nazre and Rahul Garg, "A Deep Dive in the Venture Landscape of Artificial Intelligence and Machine Learning", September 2015 (https://medium.com/@ajitnazre/a-deep-dive-in-the-venture-landscape-of-ai-ajit-nazre-rahul-garg-de0543231eb2).

Combinations of these techniques have already shaken entire sectors, starting with industrial applications (e.g. for predictive maintenance) and online platforms, from e-commerce to online search, the collaborative economy and interactive online advertising. A good example is Google search, which introduced innumerable new AI-enabled functions in its first 20 years of existence (Renda 2015). Similarly, Netflix today reportedly earns as much as $1 billion thanks to its recommendation engine, which shows users movies they could be interested in, based on previous choices (Gomez-Uribe and Hunt 2016). And Amazon invests enormous amounts of money in AI R&D to sharpen its business model and provide a more effective service to its customers. Apple, Amazon, Microsoft and Google also compete for the future of search, through their vocal assistants that make the most of recent breakthroughs in natural language processing. All these developments are shaping a world in which information is potentially easier to find, cheaper and more abundant. In an information-rich society, as Herbert Simon (1971) used to say, a "wealth of information creates a poverty of attention": AI can help us navigate through this over-abundance of information, leading us to find what is most relevant.

That said, **use cases are quickly emerging in many specific sectors, beyond the Internet economy.** These include *autonomous transportation*, which will soon be commonplace and, as most people's first experience with physically-embodied AI systems, will strongly influence the public's perception of AI; *home/service robots*, which have already entered people's homes, primarily in the form of vacuum cleaners such as Roomba; *healthcare*, where there has been an immense forward leap in collecting useful data from personal monitoring devices and mobile apps, from electronic health records in clinical settings and, to a lesser extent, from surgical robots designed to assist with medical procedures and service robots supporting hospital operations; *entertainment*, with a huge industry investing in new exciting interactive videogame experiences; and *education*, with considerable progress expected in online learning, conversational chatbots and interactive machine tutors. AI can also potentially help development and cooperation by empowering low-resource communities, and by enabling more effective policing and, more generally, public safety.

Box 2. A Knowledge Map of AI

Another way of mapping AI developments is provided by Francesco Corea (2018), who identifies a number of AI paradigms: *logic-based tools*, used for knowledge representation and problem-solving; *knowledge-based tools*, on ontologies and huge databases of notions, information, and rules; *probabilistic methods*, i.e. tools that allow agents to act in incomplete information scenarios; *machine learning*, which allows computers to learn from data; *embodied intelligence*, an engineering toolbox that assumes that a body (or at least a partial set of functions such as movement, perception, interaction and visualization) is required for higher intelligence; and *search and optimization,* i.e. tools that allow intelligent search with many possible solutions. He classifies these six paradigms into three different macro-approaches: the *symbolic approach*, which states that human intelligence could be reduced to symbol manipulation; the *sub-symbolic approach*, in which no specific representation of knowledge is provided *ex ante*; and the *statistical approach*, based on mathematical tools to solve specific sub-problems.

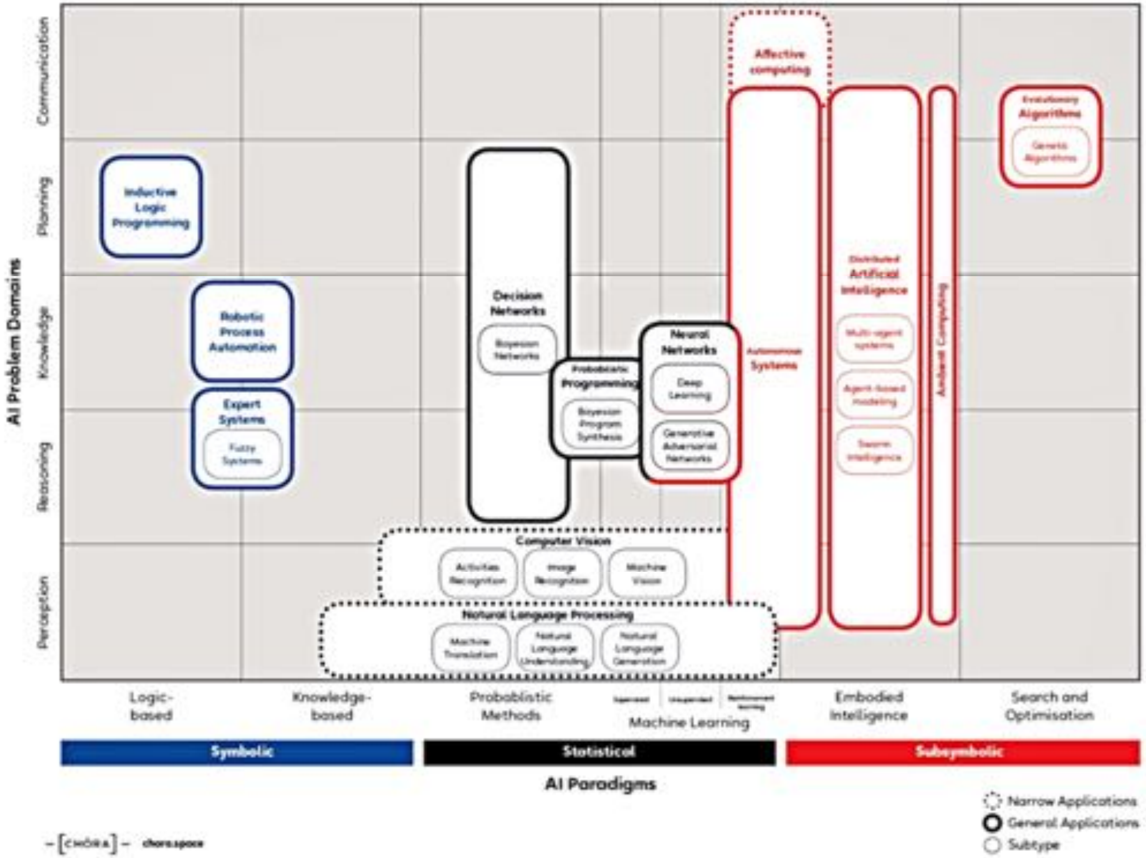
Figure 5 (on the next page) presents Corea's mapping of AI knowledge, with the vertical axis representing the problems AI is used for. A distinction is drawn between reasoning (the capability to solve problems), knowledge (ability to represent and understand the world), planning (capability of setting and achieving goals), communication (ability to understand language and communicate) and perception (ability to transform raw sensorial inputs such as images or sounds into usable information). The patterns of the boxes divide the technologies into two groups, i.e., narrow applications and general applications. Rather than hinting at Artificial General Intelligence, which remains a pure speculation today, the difference refers to technologies that can only solve a specific task (narrow applications) and others that solve multiple tasks today or in the future and interact with the world (better than many humans — general applications).

1.2 Winter is not coming (at least for AI)

While there seems to be widespread consensus that AI has the potential to revolutionize the economy, many consider the current hype to be exaggerated. In the past, artificial intelligence inspired many moments of enthusiasm and optimism, which were then followed by an "AI winter" due to lack of practical successes and a gradual loss of momentum.[9] Some commentators also mention other factors that may jeopardise the rise of AI as announced by the more enthusiastic commentators: such factors include the possible slowdown of Moore's law, which would deprive AI of the computing capacity needed to develop its most promising applications; and also the possible negative impact of AI on employment and equality, which would lead to a lack of market demand if not corrected by policy interventions.

[9] The onset of the AI winter can be traced to the US government's decision to pull back on AI research. The decisions were often attributed to a couple of infamous reports, specifically the Automatic Language Processing Advisory Committee (ALPAC) report by US government in 1966 and the Lighthill report for the British government in 1973.

Figure 5. AI Knowledge Map



Source: Courtesy of Francesco Corea (published on Forbes in August 2018).

This time round, however, **there are several reasons to believe that the AI winter will not come**. First, concerns on the end of Moore's Law appear exaggerated at best. Simply counting the number of transistors in integrated circuits does not capture the architecture and performance of modern computer processors, e.g. GPUs (graphics processing units) or TPUs (tensor processing units). Rather than focusing strictly on increasing transistor counts and clock speeds, companies now focus on performance, including power efficiency and component integration. The explosion of specialised processors for handling AI and deep-learning workloads is partly a reaction to the fact that CPUs (central processing units) do not scale the way they used to. Moreover, the current trend in designing processors is to move away from general-purpose machines to the tailoring of machines to specific applications, such as graphics and machine learning. Today, CPUs co-exist with GPUs (which improve performance by a factor of 10 over CPUs, (see i.a. Chen et al. 2013) and TPUs (which improve performance by a factor of at least 10 over GPUs[10]). CPUs perform the main tasks, GPUs do the graphics, TPUs the AI. in a related vein, the emerging trend in IT is 'parallel computing', which achieves exponential growth in throughput by using a multitude of processors at the same time, regardless of the fact that the growth of transistors in integrated circuits is slowing down. The bottom line is that even if Moore's law slows down, computing will continue to progress at a very fast pace, thanks to parallel computing, neural network structures and quantum technologies. As Moore's law becomes obsolete, technologies will find new ways to support the growth of applications, content and other hardware.

Together with stronger computation capacity, the amount of data available today is immensely greater than that available in the 1980s. Before the mid-1990s, during the pre-internet age, the availability of large datasets was an insurmountable problem: today, the volume of data available for analytics reportedly doubles each year. In addition, algorithms have significantly improved. And both public and private investment has reached unprecedented levels, in particular in the United States and China.

Participants in the CEPS Task Force agreed that AI can lead to important progress in our society, and the narrative around that view should therefore convey hope and optimism, rather than dystopian fears. As a matter of fact, as noted by Brynjolfsson, Rock and Syverson (2017), the most impressive capabilities of AI, particularly those based on machine learning, have not yet spread widely. More importantly, like other general-purpose technologies, their full effects will not be realised until waves of complementary innovations are developed and implemented. The discussion around recent patterns in

[10] https://www.zdnet.com/article/tpu-is-15x-to-30x-faster-than-gpus-and-cpus-google-says/

aggregate productivity growth highlights a similar contradiction. On the one hand, there are astonishing examples of potentially transformative new technologies that could greatly increase productivity and economic welfare (see Brynjolfsson and McAfee, 2014). There are some early concrete signs of the promise of these technologies, recent leaps in AI performance being the most prominent example. At the same time, however, measured productivity growth over the past decade has slowed significantly. This deceleration is large, having cut productivity growth by one-half or more in the decade preceding the slowdown. It is also widespread, having occurred throughout the OECD and, more recently, among many large emerging economies as well (Syverson, 2017).

Recently, **several papers analysing the impact of automation in Europe mostly find a positive contribution of robots to productivity**. Among others, Graetz and Michaels (2017, 2018) use the industrial robots database and estimate that in the 17 countries of their sample, the increased use of robots per hour worked from 1993-2007 raised the annual growth of labour productivity by about 0.37 percentage points.[11] By considering an industry-country panel specification, they found that robots appear to reduce the share of hours worked by low-skilled workers relative to middle-skilled and high-skilled workers; they do not polarise the labour market, but appear to hurt the relative position of low-skilled workers rather than middle-skilled ones. Nevertheless, the use of robots per hour worked appears to boost total factor productivity and average wages. Chiacchio et al. (2018) find that the use of robots per hour worked appears to boost total factor productivity (TFP) and average wages: however, they also find that the displacement effect (labour to capital) offsets the productivity effect, leading to job losses (see section 4.6.1 below).

Recent reports by Accenture/Frontier Economics (Purdy and Daugherty 2017), McKinsey (2017) and PWC (2017) conclude that **AI will be a game changer for total factor productivity and growth**, by gradually rising as a third pillar of production, together with labour and capital. Other research has shown similar, although often less optimistic, predictions. For example, Chen et al. (2016) estimate the cumulative economic impact of AI from 2016 to 2026 as lying between $1.5 and $3 trillion (representing 0.15 to 0.3% of global GDP). Furman and Seamans (2018) review some of the most interesting literature on the impact of AI on the economy, which mostly finds that AI and robotics have the potential to increase productivity, but they may have mixed effects on labour, particularly in the short run. They also conclude that many economists believe that "AI and

[11] The authors use data from the International Federation of Robotics and EUKLEMS to estimate robot density (the stock of robots per million hours worked) in 14 industries in 17 countries from 1993-2007.

other forms of advanced automation, including robots and sensors, can be thought of as a general purpose technology (GPT) that enable lots of follow-on innovation that ultimately leads to productivity growth".[12] The fact that AI has not (yet) translated into large productivity gains, according to Brynjolfsson, Rock and Syverson (2017), is due to a "lag between technological progress and the commercialization of new innovative ideas building on this progress which often rely on complementary investments": such a lag, these authors claim, is particularly notable in the case of GPT.

This, of course, does not mean that AI is guaranteed to succeed overnight. The CEPS Task Force members largely agreed that AI will emerge as a gradual process. As a general-purpose family of technologies, **AI will pervade all sectors of the economy and all aspects of professional and daily life. At the same time, it will have to be used responsibly: if not, winter will come for society, not for AI.**

1.3 Handle with care, not with fear

Beyond its outstanding promise, many commentators also argue that AI, if badly governed, can represent an existential risk for our society; whereas others observed that AI can make catastrophic events such as a nuclear war more likely (Geist and Lohn 2018). While **this threatening narrative should not overshadow the positive disruption that AI will bring to our society**, it is important to **map possible risks**, which will be as essential as opportunities in forming the basis for future AI policy and governance. Below, we distinguish between intentional, pernicious use of AI and unintentional damage caused by the use of AI.

1.3.1 Malicious uses of AI

Some of the emerging risks caused by the malicious use of artificial intelligence appear as the **natural continuation of existing trends**. Fake news will become "deep fakes" facilitated by Generative Adversarial Networks (GANs); phishing scams will become more sophisticated; and AI-enabled cyberattacks may become more difficult to anticipate due to the enhanced use of (unsupervised) machine learning (Renda 2018c). An authoritative report collectively published by several institutes in February 2018 argued that "the costs of attacks may be lowered by the scalable use of AI systems to complete tasks that would ordinarily require human labour, intelligence and expertise. A natural effect would be to expand the set of actors who can carry out particular attacks, the

[12] Quoting Cockburn, Henderson, and Stern (2017).

rate at which they can carry out these attacks and the set of potential targets”.[13] Max Tegmark’s 2017 book *Life 3.0* notes the concern of UC Berkeley computer scientist Stuart Russell, who worries that the biggest winners from an AI arms race would be “small rogue states and non-state actors such as terrorists” who can access these weapons through the black market. Tegmark further writes that after they are “mass-produced, small AI-powered killer drones are likely to cost little more than a smartphone”. Would-be assassins could simply “upload their target’s photo and address into the killer drone: it can then fly to the destination, identify and eliminate the person, and self-destruct to ensure that nobody knows who was responsible”.

These risks are already sufficient to generate reactions from the government side, such as the restructuring of cybersecurity and cyber-resilience plans, with the creation of pervasive, diffuse networks of data collection points, coupled with the centralization of processing power into high performance computers. However, **new risks will also emerge**. For example, the explosion in the number of connected devices and progress on miniaturization will lead to possible body hacking, which may concentrate on wearables and implants. Evidence brought to court could be manipulated just as news stories are in so-called “deep fakes”, which will have implications for parties in trials, insurers (possibly leading to higher premiums) and the outcome of litigation.[14] And the use of self-driving cars may make road traffic a favourite target for cyber-attackers; and so-called “swarming attacks” by distributed networks of autonomous robotic systems cooperating at machine speed will become possible. At the same time, as a dual-use technology, AI is also a response to other emerging risks, such as pandemics and bioterrorism. For example, companies like AIME (AI in Medical Epidemiology) have created a Dengue Outbreak Prediction platform; and scientists in South Korea have been able to train AI to detect the presence of anthrax at high speeds.[15]

Ongoing calls for a widespread global governance agreement on the use of AI and related standards appears likely to be frustrated by the emerging AI race. Certain uses of AI, however, could be subject to a global moratorium or outright ban, in order to prevent countries from engaging in a dangerous

[13] https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/1c6q2kc4v_50335.pdf

[14] For more on this issue, see the work by the European Commission on Connected and Autonomous Mobility (https://ec.europa.eu/digital-single-market/en/connected-and-automated-mobility-europe).

[15] See https://www.newsweek.com/ai-vs-bioterrorism-artificial-intelligence-trained-detect-anthrax-scientists-647856

competitive race, with possible destructive consequences for all. This is a possible outcome for the ongoing discussions on banning autonomous weapons (see also section 4.2.2 below). Academics like Toby Walsh and initiatives like the Campaign to Stop Killer Robots have denounced the escalation of this potentially destructive race, with prototype autonomous weapons under development "in every theatre of war – in the air, on the sea, under the sea and on the land".[16] Even in this case, however, difficulties in reaching agreement over the definition of autonomous weapons, and on patterns of attribution in case of distributed (e.g. swarming) attacks may lead the proposed agreement to collapse. Accordingly, while there would certainly be room for fruitful agreement in the international community, the chances that such an agreement will end up being comprehensive and effectively implemented are tiny.

At the domestic level, stability will be undermined by two other trends. The first is the threat of massive automation of work, which risks leaving parts of the population in a permanent state of unemployment or lacking the needed skills to re-enter the job market. The Global Commission on the Future of Work set up by the International Labour Organisation (ILO 2019) recently called for a universal labour guarantee, a universal entitlement to lifelong learning and the establishment of an effective lifelong learning system to face this threat.[17] Independently of what the net impact of automation will be in the end (all sorts of predictions are being made), an insistence that there will be no disruption is either naïve or false.

Governments initially tried to address this issue by considering increased reliance on universal basic income schemes, "robo-taxes" or similar policies.[18] But it is unclear whether individual well-being (typically fostered by the fact of being employed, not just by economic security, see Stam et al. 2015) and social cohesion will be materially helped by these initiatives. Accordingly, the debate has gradually moved on to a reflection on how to reskill the workforce. Inequality will increasingly go beyond the simple availability of money: unequal access to education, to political life, to high-quality services will create a risk of political disruption at home. The enormous potential of AI to reduce the cost of delivering public services may lead the more disadvantaged parts of the population to be served by "junk AI". For example, for poorer people, cheap

[16] https://www.theguardian.com/technology/2018/apr/09/killer-robots-pressure-builds-for-ban-as-governments-meet

[17] For an analysis of the jobs that will be replaced and the skills that will be needed in the future, see Accenture (2018), An Inclusive Future of Work: A Call to Action, at https://www.accenture.com/t20181114T030204Z__w__/us-en/_acnmedia/PDF-90/Accenture-Inclusive-Future-Of-Work-Full-Report.pdf#zoom=50

[18] For an analysis, see Furman and Seamans (2018).

bots may replace general practitioners, small claims judges, insurance brokers, etc., with insufficient levels of accuracy and possible ensuing discrimination or exclusion. **This is why public administrations should be held to the highest standards in delivering services, including criteria of universal access and inclusive, sustainable AI** (see Part II of this report).

The second development is related to the manipulation of public opinion through so-called "deep fakes" and new forms of disinformation campaigns, which make AI a threat to democracy. More specifically, while it can be expected that AI-powered, real-time fact-checking will dilute the possibility for post-truth political narratives, the power of AI-enabled disinformation will equally increase. **Much of this prospective impact depends on the choices governments will make to enable the diffusion of responsible, ethical and trustworthy AI; on their efforts to create a global AI community and governance, with common rules; and on national policies aimed at accompanying job automation with the gradual reskilling of the workforce**. In this respect, the efforts made by the European Union, the UK, France and sub-national governments such as Québec are to be observed with cautious optimism. Needless to say, such developments will only be sustainable if coupled with measures aimed at supporting end users in distinguishing services backed by ethically aligned AI from less ethically oriented products and services.

1.3.2 Unintentional bias and discrimination

AI can also cause harm to society and individuals unintentionally, and this is the policy issue that creates the most challenging problems for policy-makers, as well as for developers. In particular, there is widespread agreement that the use of AI can create unintentional, undesirable bias, thus violating fundamental rights and/or leading to outcomes and impacts that are perceived to be unfair. However, this statement already shows how intractable the policy problem is.

Consider again the following statement:

> The use of AI can create undesirable bias, thus violating fundamental rights and/or leading to outcomes and impacts that are perceived to be unfair.

The statement contains several elements that are controversial or difficult to interpret. First, what is undesirable bias? The problem here is that our societies are already deeply biased. For example, African-Americans in the United States are much more likely to be pulled over by the police and interrogated than are Caucasians.[19] Richer people receive higher damage awards for personal injury in courts, since damages are based on foregone earnings. Women are generally paid less than men in many sectors of the economy, other conditions e.g. level of seniority, experience and evaluations being equal.[20] Training a machine with data from the real world will in most cases incorporate these societal biases. Not surprisingly, the Google search engine was accused of showing ads for executive jobs more often to what it perceives as white males, compared to African-American women. Is this Google's fault, society's fault or simply a fact of life?

As a matter of fact, while biases already exist, the use of algorithms may in some cases exacerbate bias, amplify it or create it de novo. A recent article by ProPublica compared two stories of prisoners awaiting parole, showing how machines may end up incorporating bias from the very outset.21 In much the same vein, the use of big data and predictive policing techniques in a number of cities around the world has led to concerns over racial biases (Ferguson 2017). In 2016, many commentators argued that "AI is racist", since a beauty contest that was to be decided by an algorithm, supposedly using "objective" factors such as facial symmetry and wrinkles, led to the almost total exclusion of dark-skinned contestants.[22] Similarly, problems emerged also in large tech companies, for example when Microsoft released Tay, a chatbot that quickly began using racist language and promoting neo-Nazi views on Twitter; and when Facebook eliminated human editors who had curated "trending" news stories, to discover that the algorithm immediately promoted fake and vulgar stories on news feeds.[23] See Figure 6 for an illustration of machine bias in criminal sentencing.

What makes the issue almost intractable is that there is no such thing as a neutral algorithm: and even if it was possible to generate one, a neutral algorithm would in many cases be useless, whereas "excessively" biased algorithms can be dangerous and harmful. Accordingly, it is important to define which biases are to be considered acceptable, and which are not. There is also a potential trade-off between accuracy and privacy. In some cases, more accurate algorithms can eliminate bias by not treating people on the basis of average calculations. For example, an algorithm may decide not to grant credit to an

[19] https://www.nationalgeographic.com/magazine/2018/04/the-stop-race-police-traffic/

[20] https://www.bloomberg.com/quicktake/why-women-earn-less-than-men

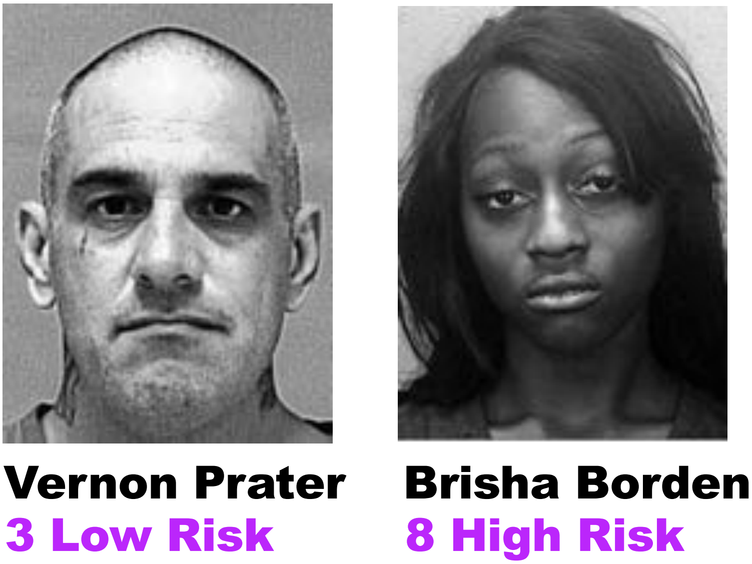
[21] Presentation by Rumman Chowdhury (Accenture) to the CEPS Task Force.

[22] https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

[23] https://www.theguardian.com/technology/2016/aug/29/facebook-fires-trending-topics-team-algorithm

individual since he or she belongs to an ethnic group that on average repays debts less often.

Figure 6. Machine bias in criminal sentencing



Source: ProPublica 2016 (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

1.4 Main findings

This section introduced the reader to the concept of artificial intelligence and its many definitions and families of methods. Sometimes confused with big data analytics or equated with machine learning, AI is in fact a much broader, constantly evolving family of methods and approaches. All in all, AI is also a general-purpose technology, which is expected to permeate most, if not all aspects of our economy, as well as our social interactions. Finally, AI is also part of a broader emerging "technology stack", which features enhanced connectivity, high-performance computing and the Internet of Things.

Given the pace of current AI development and the maturity of complementary technologies, it is fair to assume that an "AI winter" will not come this time round. That said, the development and rise of AI can represent an extremely positive force for the well-being and prosperity of mankind in the future; however, such an extremely powerful technological development can also cause intentional as well as unintentional damage. It is therefore extremely important to mitigating the possible risks, and this is why many countries, as well as corporations and civil society, are taking action to promote responsible and trustworthy uses of AI.

2. A WAY FORWARD FOR DEVELOPING AI: COMPLEMENTARITY, RESPONSIBILITY AND SUSTAINABILITY

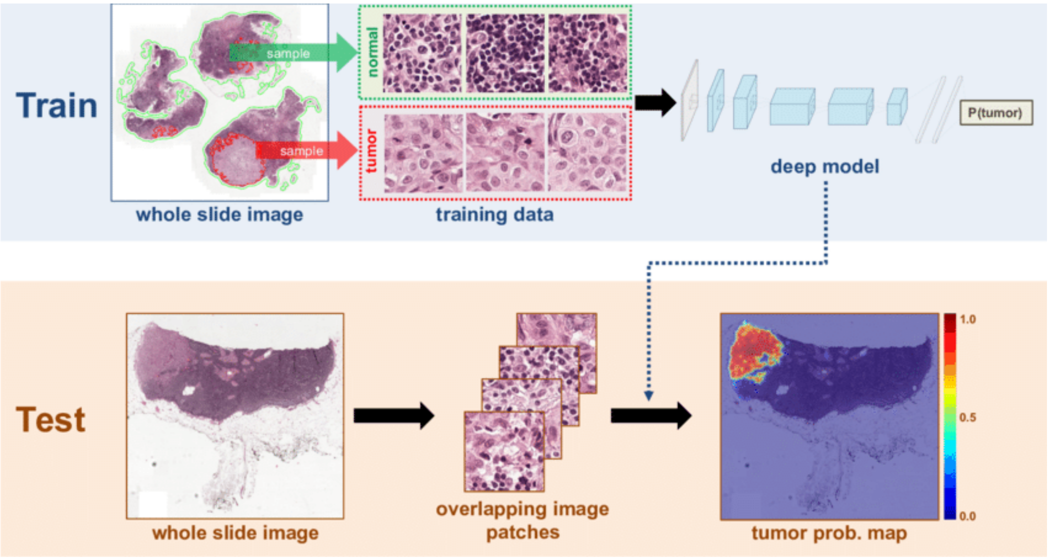
2.1 Complementarity

The first possible way to mitigate negative consequences, while at the same time harnessing the huge potential of AI is to approach it as complementary, and not as an alternative, to human intelligence. **There are several reasons to argue that AI can reach its full potential, with a minimum of associated risks, if it is coupled with a human being**. For example, trained AI has proven to be better than humans at identifying tumours from medical images: however, it often errs in very awkward ways, as shown in Figure 7, without recognising the implausibility of its results, and this requires the intervention of a human being.[24] Similarly, AI was found to significantly reduce error rates in the identification of metastatic breast cancer from sentinel lymph node biopsies, but only when coupled with an expert pathologist (see Figure 8 below).

As a matter of fact, complementarity entails that AI is used to augment human intelligence, rather than replacing it. Humans are indeed better equipped than today's trained AI machines at setting goals, using common sense and formulating value judgments; machines, on the contrary, may be better at pattern discovery, large-scale math and performing statistical reasoning. All in all, the combination of human and machine wins in most applications.

[24] Presentations by Mark Nitzberg (UC Berkeley) and Francesca Rossi (IBM) to the Task Force.

Figure 7. AI in tumour mapping

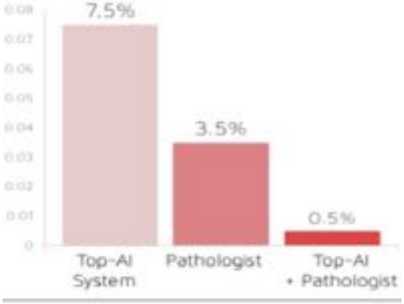


Source. June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo and Namkug Kim (2017), "Deep Learning in Medical Imaging: General Overview", *Korean Journal of Radiology*, 18. 570. 10.3348/kjr.2017.18.4.570.

A growing number of attempts to achieve full automation have backfired. One of the most famous cases is that of Tesla, whose robot-only factory producing the new Model 3 ran into trouble, as admitted by Elon Musk himself, due to the absence of human beings.[25] And the same could be said of the tragic accident that occurred in Tempe, Arizona in March 2018, when an Uber-operated Type 4 autonomous Volvo SUV killed a woman when it failed to identify her as she crossed the street due to an accidental de-activation of Lidar sensors and the lack of operation of the camera at night. The car failed to ascertain that someone (a 49-year old woman) was in its way, and the fact that the lady was walking a bike possibly presented an unprecedented situation. The unfortunate incident left the impression that self-driving cars are not yet equipped to handle all possible contingencies, let alone use common sense. The collision led Uber to suspend testing in Tempe as well as in Pittsburgh, San Francisco and Toronto.

[25] https://www.theguardian.com/technology/2018/apr/16/elon-musk-humans-robots-slow-down-tesla-model-3-production

Figure 8. AI improves error rates in cancer detection, but only working in tandem with a pathologist



AI significantly reduces pathologist error rate in the identification of metastatic breast cancer from sentinel lymph node biopsies.

Source: US White House report, "Preparing for the Future of Artificial Intelligence", 2016.

Similarly, **even when full autonomy is advertised, the underlying reality is sometimes different**. For example, the announced opening of Amazon's **check-out-free stores gives an *allure* of full autonomy, but hides the fact that,** according to a presentation hosted by the CEPS Task Force, human beings are employed in remote locations to manually correct the many mistakes due to the current inaccuracy of AI observations. Similarly, companies that claim to bring full vehicle automation to US cities indeed employ teams of individuals that remotely intervene whenever a confusing situation presents itself.[26]

More generally, **complementarity has to be accompanied by a suitable legal system**. For example, tort regimes that expose human beings to liability in the event they decide to override the decision of the algorithm are not ideally designed to make the most of man-machine augmented intelligence. A "human in the loop", knowing that no action means no liability, would not have any incentive to override the machine, even when it would patently make sense to do so. The requirement of the presence of a "responsible human" is, on the contrary, more human-centric, and likely to avoid cases in which injured individuals end up having no redress.[27]

[26] Presentation of Mark Nitzberg (UC Berkeley) to the CEPS Task Force.

[27] See, i.a., remarks by Giampiero Lotito, Founder and CEO of FacilityLive, and President of the European Tech Alliance, at the presentation of the European Internet Forum report, "Digital World in 2030", at the European Parliament in Brussels, 18 March 2014. Lotito argued against "an algorithmic society where a machine decides which information is more relevant to [end users]. This is a human capability and we must build technologies that use the human way to organize, retrieve and use information".

2.2 Responsibility and trust: Bias and value alignment

With great power comes great responsibility. Ensuring that AI develops in line with the public interest and that the enumerated risks are minimised implies that AI developers and vendors act responsibly. Responsibility can be defined in several ways and can be seen as spontaneously originating in developers and technology companies, or it can be promoted or even imposed by government. Generally, advocating responsibility implies acknowledging the potential risks of AI, and accordingly acting to mitigate them in the design, development and use of AI. Over the past two years, "responsible AI" has moved from a mere commitment to awareness-raising and educating end users to the proactive deployment of concrete tools and assets, and to the focus on enterprise applications to enable its diffusion and ultimate democratisation. Essentially, the idea of responsible AI stems from the acknowledgment of possible unintended consequences of AI development and its use, acting on essential aspects of AI such as fairness, accountability, transparency and explainability.

In his presentation to the Task Force members, Rumman Chowdhury[28] explained that unintended consequences could be tackled by addressing the following questions:

- On **fairness**: Are there factors influencing model outcomes that should not be there? For example, does the model discriminate between specific social groups or classes? Do we have an expectation of similar outcomes for different subgroups?
- In terms of **accountability**: Within a given organization, what is the chain of command for deciding what to do with a potentially biased outcome?
- In terms of **transparency**: Do we understand how the model works? And on the related aspect of **explainability**, does the model allow for identifying why and how an output was arrived at?

There are indeed many ways in which bias may creep in, when using AI-enabled algorithms. For example, the data fed into the algorithm can be biased in and of itself. This, in turn, can lead to a "garbage in, garbage out" problem, which inevitably affects the output of the algorithm, unless measures can be taken to eliminate the bias in the output *ex post* (Mittelstadt et al. 2016). Cases of **data bias** may take various forms, including selection or sampling bias and measurement bias, in which the measuring instruments or their operationalization are faulty. There may also be **response or reporting biases**,

[28] Rumman Chowdhury, Global Lead Responsible AI, Accenture Applied Intelligence.

which depend on how data are being collected, and on whether the data used are sensitive in nature, and as such likely to misrepresent the truth. It is also important to check whether the individuals that will report the data have the same metrics of reporting (e.g. so-called "yelp effect"[29]).

Moreover, it is important that designers of experiments take into account possible **design biases**, by checking the assumptions made in designing the model and its applicability to the overarching research question. Also, it is important to avoid engineering feedback loops, which occur when the results of the model are somehow fed back into the model (sometimes intentionally, sometimes not). And it is equally important to prevent players from "gaming" the system by exploiting feedback loops. For example, students who know that the system considers the fact that students who have taken advanced calculus have an increased chance of performing well in college, will take calculus simply to increase their chances of being admitted to college. Avoiding feedback loops that allow gaming the system may require periodic retraining or retooling of the system itself.

Another category of bias that should be brought to the attention of AI developers is **societal bias**, which is reflected and exacerbated by algorithms (Future of Privacy Forum 2017). For example, filtering candidates by work proximity can lead to economic loss: in 2012, Xerox reportedly caught a possible bias in a recruiting algorithm before implementing it: even though proximity was a sign of retention, it also excluded minorities who tended not to live close to Xerox due to housing prices. Economic loss due to the narrowing of choice emerges also when career recommendation engines constrain concept of career prospects due to historical data of what a successful path looks like: or when internal job search results are based on who you know. Mitigation strategies include understanding what proxies are used and are they directed towards protected classes; thinking through reasonable alternatives to get at the outcome; reinforcing strong data hygiene practices; building explainability and understandability into systems. Disparate impact occurs when a company employs facially neutral policies or practices that have a disproportionate adverse effect or impact on a protected class, unless those practices or policies further a legitimate business need that cannot reasonably be achieved by means that are less disparate in their impact (Federal Trade Commission 2016).

Besides bias, another area that is closely linked to responsible and trustworthy AI is the so-called **"value alignment" problem**, according to which autonomous AI systems should be designed to ensure that their goals and behaviour can be aligned with human values throughout their operation. There was consensus in the presentations hosted by the CEPS Task Force on the fact

[29] https://hbswk.hbs.edu/item/the-yelp-factor-are-consumer-reviews-good-for-business

that **value alignment should also apply to current AI systems, not just to future, highly autonomous systems, as the corresponding Asilomar principle suggests.**[30] At the same time, ensuring value alignment is far from easy, for many reasons.

2.2.1 *Which values should AI be aligned with?*

Recent research by Freedman et al. (2018) on kidney exchanges and research by Awad et al. (2018) on the MIT Moral Machine, which simulates the trolley problem on self-driving cars, has shown stark differences in the decision-making criteria that individuals would consider acceptable in case of "life or death" decisions. The scholarly field of ethics provides an interesting framework, but certainly no ultimate conclusion. For example, machines could be trained to be *utilitarian* (Bentham, Mill), and thus focus on the ultimate result of their action and the rationality of their behaviour. Alternatively, a deontology-focused approach (Kantian) would focus on the law, as well as on moral imperatives and actions that are considered to be ethical or unethical, regardless of the result. Finally, virtue ethics, as explained by Dignum et al. (2018) and by Berberich and Diepold (2018) and rooted in the work of Aristotle, focus on motives and are *relational* rather than rational, in that they focus on following virtuous examples. **Deontology and virtue ethics focus on individual decision-makers, while teleology considers all affected parties. None of these approaches provides uncontroversial, definitive ways to resolve conflicts.**

2.2.2 *How can value alignment be achieved from a technical standpoint?*

Scholars and experts have listed several activities that can be said to broadly belong to value alignment. In her presentation to the Task Force, for example, Francesca Rossi pointed out different techniques that can lead to value alignment in reward systems, including reward-based personalization policies, which can however lead to unethical recommendations; exogenous ethical policies that cannot be modified by user's response to recommendations, which are learnt offline and cannot be modified during online usage of the system; or combinations of **ethical and reward policies**. Other authors point to the differences between **imitation learning, apprenticeship learning and inverse reinforcement learning**. The latter is proposed by Russell, Dewey and Tegmark (2015) as a particular machine-learning approach to ethically training autonomous systems: in this technique, instead of rendering rules, laws or utilities from the start, the system learns from modelled behaviour what an actor is trying to do and what kinds of behaviour are being sought (Arnold, Kasenberg and Schutz 2017).

[30] https://futureoflife.org/ai-principles/

More generally, as observed by Virginia Dignum, the challenge of merging AI with ethics can be solved through a holistic approach, aimed at promoting ethics "in" design, "by" design and "for" designers. More specifically:

- **Ethics "in" design** requires ensuring that development processes are aligned with ethical principles. This, in turn, implies that the ethical implications are taken into account as AI permeates our society, integrating and replacing traditional systems and social structures. This branch of ethics in AI is closely linked to accountability, transparency and responsibility in designing AI. It increasingly makes use of data science concepts, especially when it comes to embedding fairness considerations in a step-by-step analysis of the AI development process. Scholars from several disciplines have contributed to the conceptualization of "quantitative fairness" (Hutchison and Mitchell 2019), a process that resulted in a variety of possible definitions (Verma and Rubin 2018; Naranyan 2018). These metrics are now being embedded in machine learning (Barocas et al. 2018) with the aim of tracing the causes of bias in machine learning, e.g. skewed samples (initial bias that compounds over time), "tainted examples" (basing the machine learning process on a non-representative example), limited or wrongly chosen features and sample size disparity or misuse of proxies (Barocas and Selbst 2015). The observance of ethical principles "in design" is facilitated by emerging interactive software tools such as Accenture's Algorithmic Fairness Tool (Chowdhury 2018); and IBM's AI Fairness 360 toolkit, Factsheets for AI services and support for Supplier Declarations of Conformity (Hind et al. 2018).
- **Ethics "by" design** requires the integration of ethical reasoning abilities as part of the behaviour of artificial autonomous systems (such as agents and robots). This domain includes "methods, algorithms and tools needed to endow autonomous agents with the capability to reason about the ethical aspects of their decisions, and the methods, tools and formalisms to guarantee that an agent's behaviour remains within given moral bounds" (Dignum et al. 2018). Such techniques include Inverse Reinforcement Learning, Apprenticeship and others. The limitation of (relying exclusively on) this approach is that it is still uncertain whether we will ever succeed in building fully ethically aware agents, able to tailor their decisions to the context in which they operate.
- **Ethics "for" design(ers)** refers to the need for integrity of researchers and manufacturers as they design, construct, use and manage AI systems. This approach to AI ethics implies reliance on deontological approaches (e.g. Codes of Conduct), which may also come with

enforcement possibilities (e.g. algorithmic inspections and verification, certification and mandatory explainability).

It clearly appears that none of these approaches can and should be seen in isolation; a sound policy for ethically aligned AI is likely to rely on a combination of these approaches and should be designed with due attention to the principle of proportionality. Several attempts have been made at building a comprehensive framework for responsible, trustworthy AI so far. Some of them limit themselves to declarations of principles (e.g. Asilomar principles), whereas others effectively provide a framework for implementing AI and aligning it with agreed-upon values. None of them provides a final answer to the achievement of responsible AI.

2.3 Sustainability

A key element of future AI development is sustainability, intended in particular from an economic, social and environmental perspective.

Social sustainability is particularly important when it comes to AI. Looking at the UN's Sustainable Development Goals (SDGs), several social elements are potentially affected by AI development. In particular, the SDGs pursue full and decent employment and enhancement of human capital as key objectives, which seems to be hardly consistent with evidence of accelerating job replacement due to automation. Also, regardless of whether jobs will be replaced, goals related to eradicating hunger and poverty and reducing inequality inevitably constrain AI development, for good.[31]

Environmental sustainability is an often-neglected aspect of AI development, and can be approached from several angles. One of them relates to AI's carbon footprint, which seems to be controversial. The global energy consumption of data centres has been estimated at 194 TWh in 2014, which is around 1% of annual global electricity consumption, i.e. more than the electricity consumption of several EU member states. Data centre consolidation, outsourcing and cloud computing are helping to keep energy consumption in data centres flat, notwithstanding the increase of data and processing, as larger data centres tend to be more efficiently designed and managed. The solution to this problem seems to be rooted in technological developments, and in particular in AI. GPUs, TPUs, new protocols and AI solutions can dramatically improve energy efficiency.

[31] "First, to ensure that the development of AI technology does not cause an increase in social and economic inequality. Second to call on AI in order to reduce this" (Mission Villani, 2018: 133).

The sustainability dimension has been captured in particular by the "AI for good" initiative, championed by the International Telecommunication Union (ITU) and more generally by the United Nations. The Mission Villani report also set the goal in France to develop AI as a tool for a sustainable and ecological economy, providing a vision for a "greener" AI enabling an ecological transition (Mission Villani 2018). And a recent report for the World Economic Forum, co-authored by PwC and the Woods Institute at Stanford, advocates AI for the Earth as a form of Value Alignment.

2.4 Conclusion: An enormous potential, in need of direction

We are increasingly discovering the potential of AI – and its risks. Both are huge. The more we see AI getting out of the labs and permeating society, the more we need to get ready to make it compatible with our values and our needs. In this respect, it is AI that must be adapted to our legal systems, rather than the other way around. In some cases, however, the legal system will need to be changed to contemplate new, AI-enabled ways of providing goods and services, organising production and social interaction.

In this section, we have identified three main directions in which AI should move in order to remain aligned with the interests of mankind: complementary, responsibility and sustainability. All three require sustained attention by the public as well as the private sector and are being targeted by various initiatives, declarations and manifestos around the world. The next step is to make them actionable in terms of policy.

PART II. EUROPE IN THE GLOBAL RACE FOR AI

The European Commission has stated its intention to lead the way in developing and using AI "for good and for all" and is taking concerted action to this end. Good intentions alone, however, are not sufficient to achieve the intended results, and many scholars and experts have expressed scepticism about Europe's ability to set rules and actively compete in the AI domain. Such doubts are related to both prongs of Europe's current AI strategy: its ability to be a global norm leader and standard-setter, which possibly stems from the recognised strength of its legal system and emphasis on values and norms; and its alleged leadership in academic research as well as in specific industrial applications. We briefly summarise below the current global competition for AI, and Europe's current efforts in terms of spending programmes, law-making and multi-stakeholder consensus-building. We then turn to the two main fronts on which the European Commission is focusing, with the help of the AI High-Level Expert Group (AI HLEG)[32] and the European AI Alliance[33]: i) drafting ethical guidelines on AI and ii) defining a policy and innovation strategy to build competitiveness in this key domain.

[32] See https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

[33] See https://ec.europa.eu/digital-single-market/en/european-ai-alliance

3. GLOBAL AI COMPETITION: WHY GLOBAL STANDARDS WILL NOT EMERGE WITHOUT CONCERTED ACTION

Governments at all levels of development have realized that AI, and more generally the whole emerging "technology stack" (as defined in section 1 above) can have important repercussions on the global order and its balance of powers. At the lower layer of the ecosystem, the race for supercomputers is leading to important breakthroughs, such as the evolution of computing capacity into parallel computing, in which different processing units perform different functions (CPUs the main tasks, GPUs the graphics and TPUs machine learning and other forms of AI). An even bigger breakthrough is expected when quantum computers will reach a suitable capacity (i.e. number of qubits).[34] Quantum and other forms of chips (e.g. biological, neuromorphic) are expected to reach the capacity of the human brain (approximately 85 billion neurons) by 2025, and then skyrocket afterwards to unknown frontiers. The country that wins this race will achieve key advantages, especially in cryptography, with applications mostly in scientific research, complex optimization issues and, inevitably, cyberwarfare.[35]

The geopolitical relevance of this race cannot be overstated. Suffice it to recall what Vladimir Putin recently said in a speech: AI "is the future, not only for Russia but for all humankind. Whoever becomes the leader in this sphere will become the ruler of the world".[36]

[34] Google's recent decision to open-source its quantum computing platform Cirq is a good demonstration of the importance of securing competitive advantage early on in this rising domain. (See https://ai.googleblog.com/2018/07/announcing-cirq-open-source-framework.html).

[35] It should therefore come as no surprise that China has intensified its efforts to file patent applications for quantum cryptography, an area in which it massively dwarfs other countries today. See "Quantum technology is beginning to come into its own", The Economist Technology Quarterly – Here, There and Everywhere, 2017 (https://www.economist.com/news/essays/21717782-quantum-technology-beginning-come-its-own) and Purdy and Dougherty (2017).

[36] https://www.youtube.com/watch?v=lHd7s3I3Zb4

Indeed, **China has stated its ambition to become the global leader in AI innovation by 2030**. The superpower is coupling very ambitious plans on AI with a rising position in research and innovation in all aspects of the technology stack, including in particular supercomputing, 5G connectivity and IoT development. President Xi Jinping (2014) publicly stated that China "should unswervingly follow an independent innovation path featuring Chinese characteristics, stick to the guiding principles of independent innovation, leap-frogging development in key sectors, with development supported by science and technology and oriented towards the future". Over the past five years, the country has launched several initiatives with a view to realising a €13 billion AI market in China by 2018 and making China the world leader in AI by 2030. These include the "Made in China 2025" campaign launched in 2015, which focused on precision and agile manufacturing; the "Internet +" Action focused on smart manufacturing and innovation; the Robot Industry Development Plan (2016-2020) launched in 2016 to foster the development of intelligent industrial and service robots; and most notably the "New Generation AI Development Plan" launched in 2017 to achieve well defined milestones such as placing China on a par with other leading economies by 2020; **attaining key breakthroughs in AI theory by 2025 and start conquering the world market with them; and ultimately becoming the world leader in 2030.** The Chinese plan is sharp and complete, covering both military and civilian applications of AI, and is fraught with ubiquitous technological developments, which bring the promise of an extremely efficient society, powered by massive and pervasive use of technology. Feijoo et al. (2019), as reported in JRC (2018), consider the Chinese plan to be ambitious but achievable, and their judgment is supported by other analysts.

On the other hand, **the United States seem to be adopting a much more "hands-off" approach to Artificial Intelligence**, also thanks to the large R&D expenditure of its tech giants and leading universities. In May 2018, the White House announced its broad intention to maintain American leadership in AI, support the American worker, promote public R&D and remove barriers to innovation (JRC 2018). More recently, the US government's Defense Advanced Research Projects Agency (DARPA) announced the so-called "AI Next" programme, a $2 billion investment plan aimed at addressing the perceived limitations of current AI technologies, including excessive data-dependency, lack of explainability and lack of contextual reasoning. But the US federal government as a whole does not seem likely to adopt a strategy for responsible AI any time soon.

Other countries that have adopted dedicated AI strategies include **Japan (2015), South Korea (2016), Canada (2017) and India (2018)**. Among these countries, Japan seems the most advanced in proposing the integration of AI and robotics with society, in what the government calls "Society 5.0". In 2017 Japan published a new strategy aimed at increasing productivity through user-driven hyper-customisation of services, medical health care and welfare to support an increasingly ageing population, mobility to support safe and environmentally friendly travel for all, and information security (JRC 2018). This strategy comes with a pathway to 2030, built on four steps and on three main centres of excellence in charge of complementary actions.

Such bold initiatives have triggered plans in many European countries, including the UK, Finland, France, Sweden and most recently Germany.[37] These and other national strategies have motivated the European Union to take action, in part to avoid the implementation of widely diverging national plans, which would most likely be counter-productive given the size of Europe's global competitors. The EU institutions have also worked to promote more coordination between member states through the so-called "Digitising European Industry" group, which contributed to the EU coordinated plan published in December 2018 (see below).

All in all, while China and the US appear determined to invest in industrial applications to boost their global competitiveness, Europe seems to be more concerned with the need for responsible, trustworthy AI, which falls in line with key ethical and legal principles, thus avoiding a race to the bottom in AI development. Countries like China and Russia now aim at increasing the speed of AI development by securing massive data availability for machine training: such data are often collected through very intrusive technological means, such as facial and body recognition, or even "social credit scores". With a projected one trillion connected devices by 2035, the unconstrained collection of data may lead to unpredictable changes in the way governments relate to citizens and how the latter organize their social lives. To be sure, in these countries the AI race may end up sacrificing the protection of personal data on the altar of faster, more capable machines: an "unintended" outcome that some of these governments may not regret after all. The risks are inevitably high: not surprisingly, AI was recently portrayed as likely to increase the likelihood of a nuclear war in the coming two decades (Geist and Lohn 2018). Making the right choices is indeed essential for the harmonious development of AI at the domestic and international level. We address specific issues below, related to both domestic and international aspects.

[37] The European Commission reported in its Coordinated Plan for AI, adopted on 11 December 2018, that Denmark, Luxembourg, the Netherlands, Ireland and Norway include AI-related actions in their broader digitization strategies; whereas Austria, Belgium, the Czech Republic, Denmark, Estonia, Germany, Italy, Latvia, Poland, Portugal, Slovenia, Slovakia and Spain are in the process of developing strategies.

3.1 Where is AI going? Between global good and the temptation of sovereignty

It is difficult to disentangle the emerging trends in global governance from the additional effects that will be generated by the evolution and diffusion of AI technologies. Already now, China is on its way to becoming the most powerful global economy, with the United States currently stepping back from the proactive, almost uncontested leadership that it has enjoyed over the past few decades. This is reflected in many domains of global governance, including climate policy, trade policy and to some extent even the G7/G20. China is rising as a would-be global leader even in environmental policy, and possibly also as a contributor to global peace and stability (at least in the neighbouring regions), and as a would-be leader in general-purpose new technological developments such as AI. This, coupled with the booming demography of countries like India and Nigeria, provides a first idea of the terrain on which AI will develop.

The international political order will be heavily affected by this transition. In particular, it is reasonable to expect that due to the data-hungry nature of current AI applications (mostly based on machine learning), and the pervasive nature of such applications, the emerging technology stack will be considered as "critical infrastructure", i.e. essential to national stability in the near future, most likely within the next ten years in many developed countries. The explosion of the Internet of Things and the massive generation of data-driven, AI-powered applications that run key critical infrastructure such as energy grids, internet pipelines, the food chain, the ATM network, hospital logistics and care delivery will gradually lead countries to try to protect the IT stack as a domestic asset. The risk of foreign "intrusion" into the data architecture, already existing today (suffice it to think about Russia's meddling in US elections), will gradually become an existential risk for governments. Thus, **a temptation to invoke so-called "AI sovereignty" or "AI autarchy" may emerge**, just like sovereignty-related sentiments and reactions were elicited by the Snowden revelation related to NSA mass surveillance, especially in countries like Germany and France; and the threat of Russian or Chinese meddling into elections triggered reactions in the US, Italy, the UK and also recently in Australia.[38] "AI sovereignty" will be even more loudly invoked in the age of quantum supremacy, given the need to avoid that advances in cryptography provide hostile nations with important strategic advantages in global intelligence. Despite the inherently global nature of technologies like the

[38] See i.a. C. Hamilton, "Australia's Fight Against Chinese Political Interference", *Foreign Affairs*, 26 July 2018.

internet and AI, such a tendency may emerge both in non-democratic countries and in democratic ones.

The most obvious response to this potential trend would be to develop deeper cooperation on the relationship between **AI and international human rights**. In this respect, the Toronto Declaration prepared by Amnesty international and Access Now proposed a framework for reconciling AI development with the International Human Rights framework. However, despite alarming findings on the misuse of AI and, more generally, big data analytics as "weapons of math destruction" (O'Neil 2016), or as tools liable to "automate inequality" and exacerbate lack of accessibility; and despite the mounting evidence of use of AI systems to manipulate public opinion and meddle through domestic elections, let alone score and rank citizens based on very intrusive personal data mining, the **global community does not seem likely to reach an agreement on minimum standards of responsible use of AI: the stakes are simply too high**. This also implies that autonomous weapons, cyberwarfare and possible negative effects of AI on jobs, social equality and cohesion, and the environment may not be subject to a global governance effort in the next few years.

A different angle to the interface between global governance and AI, which might hold more promise in the international community, is the **incorporation of AI in the overall discussion on the Sustainable Development Goals**. This approach, represented in ongoing initiatives such as the ITU's "AI for Good Global Summit", focuses on the uses of AI that can help the global community achieve the SDGs. This focus was also shared and echoed by several large private companies and foundations, which profess their commitment to achieving the 2030 Agenda goals through enhanced use of AI. However, looking at current trends – such as the resurgence of nationalism in politics, deteriorating rule of law in some European countries, new protectionist stances and tariff wars in trade, short-termism in social policy and reiterated denial on climate change – the agreement reached in September 2015 by 193 countries on the SDGs seems to belong to a very distant era in human history. Indeed, today the United States has reached a record low in its commitment to SDGs, Brazil is entering a new era of populism and China struggles to show leadership on environmental, and even more social, achievements. Recent reports confirmed that, with the exception of Scandinavian countries, all high-income countries are far from a trajectory that would lead them to achieve the 17 SDGs and are struggling in particular with four objectives related to sustainable consumption and production patterns, climate action, aquatic life and life on land.[39]

[39] https://www.un.org/sustainabledevelopment/sustainable-development-goals/

In general, the **current landscape seems to highlight the existence of a huge gap in leadership on AI global governance: a gap that the US and China are probably unwilling to deal with, and that only the European Union, working as a collective, would have the strength to fill**. The EU could become a leading voice in a new global governance setting where technical standards are otherwise being shaped only through voluntary, multi-stakeholder, transnational private regulation (e.g. through private technical standardization bodies such as IEEE or ISO).

4. THE EU'S EMERGING STRATEGY FOR AI

For more than a decade now, the EU institutions have been adopting policy initiatives and expenditure programmes in fields related to AI. Looking at the whole technology stack of AI, it is important to recall the extensive funding of research and innovation in the domains of high-performance computing (HPC), including a decade-long flagship initiative that has led to the emergence of a vibrant research community in this field. Let us also not forget the many initiatives on the platform economy, including the proposed regulation on platform-to-business, the regulation on the free flow of data, the ongoing debate on the role of online intermediaries in countering copyright infringements, hate speech and disinformation, and of course the GDPR (General Data Processing Regulation) with its provisions on profiling and explainability. Looking more specifically at Artificial Intelligence and robotics, the EU has laid more concrete foundations of its AI policy since 2016, when the European Parliament adopted its first draft resolution on "Civil Law Rules for Robotics". That initiative, then adopted in 2017, portrayed a rather dystopian view of AI, by evoking Mary Shelley's Frankenstein on its first page, and calling for attributing both "rights and duties" to smart autonomous robots, an idea that was firmly rejected by several academics.[40] It also called on the European Commission to reflect on the creation of a possible Agency for AI in Europe, a step that the European Commission rightly found to be premature. However, despite some exaggeration and an overall negative narrative, the initiative adopted by the Parliament paved the way for a more organic approach to AI policy in the European Commission.

One year later, in the mid-term review of the Digital Single Market strategy, the European Commission highlighted the importance of securing a leading position in the development of AI technologies, platforms, and applications. In October 2017, the Council invited the Commission to put forward a European approach to artificial intelligence. Other EU institutions, such as the European Economic and Social Committee, also published

[40] See the Open Letter to the European Commission on Artificial Intelligence and Robotics, at https://g8fip1kplyr33r3krz5b97d1-wpengine.netdna-ssl.com/wp-content/uploads/2018/04/RoboticsOpenLetter.pdf

communications on Artificial Intelligence, and Member States started to develop their own strategies (EESC 2016). In April 2018, following a political agreement between 24 member states and Norway on cooperation in AI, the **European Commission Communication on "Artificial Intelligence for Europe"** saw the light.[41] The Communication, issued in parallel to the Commission Communication "Towards a common European data space", adopted a more positive narrative on AI compared to the European Parliament's initial resolution, and laid the foundations for a comprehensive AI strategy, by clarifying the main elements of the intended EU "secret sauce" on AI. The main assumption behind the strategy is that Europe "can lead the way in developing and using AI for good and for all, building on its values and its strengths", and that in so doing it can capitalise on good fundamentals, in particular world-class researchers, labs and start-ups; strength in robotics and world-leading industries (transport, healthcare, manufacturing); the Digital Single Market; and a "wealth of industrial, research and public sector data which can be unlocked to feed AI systems".[42]

The main assumption, i.e. that "Europe can lead", came with three separate, but complementary commitments: i) to increase investment up to a level that matches Europe's economic weight; ii) to leave no one behind, in particular when it comes to education and ensuring a smooth transition towards the AI age in the workplace; and iii) to base new technologies on "values". With respect to latter commitment, the Commission made explicit reference to the GDPR, at the time still not in force, as well as to Article 2 of the Treaty on EU, which mentions explicitly "human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities"; and a "society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail".

The Communication also announced the adoption of a series of initiatives on AI, including the creation of a High-Level Expert Group on AI (AI HLEG), as well as the launch of an AI Alliance, which quickly attracted many adherents (2,656 participants had registered as of 4 February 2019). The AI HLEG, counting 52 experts, was tasked with the definition of ethical guidelines, a first draft of which was published in December 2018; as well as the formulation of policy and investment recommendations, which should see the light in mid-2019. This report is aimed i.a. at contributing to the work of the High-Level Expert Group: as such, our main findings below follow a similar structure to the one that is

[41] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions – Artificial Intelligence for Europe, COM(2018) 237 final.

[42] Ibid., p. 2.

currently being adopted by the Group. This section comments on the content and scope of the future ethical guidelines; whereas section 5 contains our analysis of the possible content of the policy and investment recommendations.

4.1 Towards ethical guidelines: The EU's "secret sauce" for AI

The world is already inundated with guidelines, manifestos, statements and lists of principles. At the EU level, the European Group on Ethics in Science and New Technologies (EGE), an independent advisory body of the President of the European Commission, produced a statement in March 2018 on "Artificial Intelligence, Robotics and 'Autonomous' Systems",[43] which laid down a number of core values for AI in Europe, elaborated below:

- *Human dignity*, understood as the recognition of the inherent human state of being worthy of respect. This implies i.a. limits to the profiling of individuals on the basis of algorithms and 'autonomous' systems, especially without their explicit consent; but it also extends to the right to be informed of the non-human nature of interfaces with which users are interacting. A similar provision was adopted at the end of September 2018 in California.[44]
- *Autonomy*, implying that AI-enabled systems must not impair the "freedom of human beings to set their own standards and norms and be able to live according to them", a principle that potentially extends to the need to preserve control over the predictability of the outcomes of algorithmic decision-making.
- *Responsibility*, in particular when it comes to the alignment of AI-enabled systems with the global social and environmental good.
- *Justice, equity, and solidarity*, which entails that AI-enabled systems contribute to global justice and equal access; and accordingly prevent or,

[43] See the EGE Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems (https://ec.europa.eu/info/sites/info/files/european_group_on_ethics_ege/ege_ai_statement_2018.pdf).

[44] It is unlawful for any person to use a bot to communicate or interact with another person in California online, with the intent to mislead the other person about its artificial identity for the purpose of knowingly deceiving the person about the content of the communication in order to encourage a purchase or sale of goods or services in a commercial transaction or to influence a vote in an election. A person using a bot shall not be liable under this section if the person discloses that it is a bot.

See https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001.

where appropriate, timely detect and/or report "discriminatory biases in data sets used to train and run AI systems".[45]

- *Democracy*, which implies that key policy decisions are taken with the strong involvement of civil society; and that AI is not used to disrupt the functioning of our political systems and manipulate public opinion.
- *Rule of law and accountability*, including the clear attribution of liability in case AI systems cause damage to third parties, or infringe human rights such as safety and privacy.
- *Security, safety, bodily and mental integrity*: This, according to the EGE group, requires a broad notion of safety, encompassing the "external" safety of AI for its environment and users; the "internal safety" in terms of reliability and robustness; and so-called "emotional safety" with respect to human-machine interaction. This principle of course extends to the use of AI to develop lethal autonomous weapons, which the European Parliament has officially proposed to ban in a recent resolution.[46]
- *Data protection and privacy*, which the EGE proposes to extend beyond privacy *stricto sensu*, to possibly encompass "the right to meaningful human contact and the right to not be profiled, measured, analysed, coached or nudged".[47]
- *Sustainability*, intended by the EGE group essentially in its environmental dimension, but possibly extended also to the economic and social dimensions.

Of course, this is not the only list. For example, Floridi et al. (2018) compare the EGE statement with five other documents: the Asilomar AI principles; the Montréal Declaration for Responsible AI; the General Principles offered in the second version of the IEEE "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems"; the five overarching principles for an AI code developed by the UK House of Lords (2018); and the Tenets of the Partnership on AI (2018). Already these documents lead to a total of 47 different principles, although with significant overlaps: if one adds that there are other documents circulating, including i.a.

[45] See the EGE statement on Artificial Intelligence, Robotics and 'Autonomous' Systems (https://ec.europa.eu/info/sites/info/files/european_group_on_ethics_ege/ege_ai_statement_2018.pdf).

[46] European Parliament resolution of 12 September 2018 on autonomous weapon systems, 2018/2752(RSP).

[47] See the EGE statement on Artificial Intelligence, Robotics and 'Autonomous' Systems (https://ec.europa.eu/info/sites/info/files/european_group_on_ethics_ege/ege_ai_statement_2018.pdf).

the "Toronto Declaration on protecting the rights to equality and non-discrimination in machine learning systems", the identification of values and principles for AI development already looks like a quagmire.

It is therefore important that the European Commission, backed by the AI High Level Expert Group, does not limit itself to the adoption of another list of principles, but develops a list that represents the EU approach to AI, and offers concrete guidance to stakeholders on the "what" and "how" of selected principles to be followed and implemented. It addresses, for example, which applications or business models are potentially problematic and which ones are to be prohibited; what are concrete implications for AI development, both overall and in different fields of application; and how could developers ensure that their practices are aligned with the guidelines. In this respect, the fact that the AI HLEG is oriented towards adopting principles that were originally developed for bioethics does not seem to be the most promising approach. On the contrary, as will be discussed in more detail below, principles of bioethics can impose excessive burdens on AI systems if applied without paying heed to the principle of proportionality or giving guidance on how principles will be endorsed or enforced (Nabi 2018).

4.1.1 To "whom" should the ethical guidelines be addressed?

The future EU Ethics Guidelines on AI should be essentially addressed to the "supply side", including AI developers, vendors, and distributors; and also to organisations using or deploying AI, and public administrations, which should decide whether and how to use AI in their daily activities, and how to procure AI products in a way that is aligned with EU values and existing legislation. In order to maximise their effectiveness and the consistency of application across users, the Guidelines should not contain separate sections for different types of actors. Rather, the Guidelines should include a definition of values and principles that developers, vendors and distributors of AI, as well as organisations deploying AI can adopt as reference when placing AI systems on the market – whether B2B (business to business), B2C (business to consumer) or G2C (government to citizen) – or using AI systems in-house to improve their back-office operations.

This, of course, does not mean that no initiative should be taken on the "demand side". Those on the receiving end should be able, whenever possible, to discern when the AI system that is being used is not compliant with the fundamental principles, and those subset of cases in which this would lead them to the possibility of seeking redress since the system in use violated their rights as protected by the EU legal system (see next section). Moreover, the European Commission should work with member states to launch awareness-raising and educational campaigns that would enable end users to become more responsible and empowered individuals when dealing with AI systems and interfaces.[48]

In this respect, while developers, distributors and vendors will be directly addressed by the guidelines, public authorities and end users should also be considered as indirect addressees of the forthcoming document and future associated initiatives. As will be specified in more detail below, this also has important consequences for the content of the guidelines, as well as their relationship with enforcement activities.

4.2 "What" should be included in the Ethical Guidelines

Based on the work of the CEPS Task Force on Artificial Intelligence, we propose that the forthcoming EU guidelines consist of four main sections containing: i) an enumeration of EU values and principles of responsible, accountable and sustainable Artificial Intelligence (REACH); ii) a listing of AI applications and use cases that appear to be problematic from the standpoint of the application of the identified principles, as well as use cases in which there is no specific ethical problem raised by the AI system; iii) guidance on possible measures that could be adopted by AI developers, vendors and distributors to check that their AI-enabled applications and systems are aligned with the ethical guidelines; and iv) users' rights and possible enforcement measures and channels for redress in case of infringement of those principles outlined in the guidelines that are mandatory based on EU law. These guidelines should be available as an online, living document, so that good and bad practices can be regularly updated and kept in line with technological developments.

4.2.1 First section: A taxonomy of principles

In terms of the **principles to be identified**, it would be important that the European Commission identifies different layers of principles, in line with our observations in section 1.4 above:

- ***Level 1. Fundamental principles (Lawful AI).*** One first definition of those principles that, being rooted in the EU core values, are in any event *mandatory* for AI developers, vendors and distributors. Some of these principles are clearly defined in existing legislation, as is the case for data protection under the GDPR. Others are less clearly interpretable by market players, especially when they are rooted in the EU Treaties: this implies that any AI system that fails to respect these principles could be brought before

[48] An initiative that moves in this direction is the AlgoAware website funded by the European Commission (https://www.algoaware.eu/).

a court and be declared unlawful. As a preliminary list, these principles should include:

- *The "non-maleficence principle" ("do no harm")*. This principle, commonly applied in bioethics and technology policy, should be subject to interpretive guidance, and be tightly related to the subjective element of "intent" behind the development of an AI system. Based on the non-maleficence principle, AI systems should not be designed with the intent to harm human beings. That said, beyond the intentional element, the fact that an AI system causes damage should not automatically imply that the system is unlawful *per se*. Much in the same vein, there should be explicit treatment of those cases in which "life or death" decisions are pre-programmed (as may be the case for future automated vehicles in case of a "trolley problem"): while these types of situations should be considered as the exception rather than the rule, it should be explicitly stated that the "life or death" context should be avoided by AI developers even if this comes at a cost (e.g. developing separate *ad hoc* infrastructure for automated vehicles, or for pedestrians; or multiplying the sources of information to avoid that the vehicle fails to spot a human being, even if this is not the most economically efficient solution) (Renda 2018a).
- *Protection of human integrity, security and privacy*. This general principle encompasses various element of the protection of the individual. In particular, the requirement to protect human integrity and security implies that AI systems are not developed without exercising special care in checking that they do not have the potential to cause harm to individuals. And the protection of privacy is rooted in the GDPR and should extend to the application of safeguards against the circulation of personally identifiable data to third parties without the explicit consent of the data subject. The GDPR also gives rise to a selected number of user rights, which should be considered as inalienable and thus part of the "level 1" principles of AI development: these include the "right not to be subject to a decision based solely on automated processing", which establishes a number of safeguards designed to ensure the "fair and transparent processing" of personal data, including an obligation that entities provide "meaningful information about the logic involved" in certain types of highly automated decision-making systems; and the so-called "right to be informed," or, most commonly, a "right to explanation." The GDPR, in its recitals, also clarifies that when dealing with personally identifiable data, companies should ensure fair and transparent processing by adopting a variety of procedural and

organisational measures, which would not otherwise be mandatory by law.[49]

- *Principle of respecting human dignity, including the right not to be discriminated against*. This entails that AI systems are not designed with embedded discrimination based on criteria such as gender, ethnicity, or membership in a minority or protected group. They should also avoid violating solidarity principles (e.g. in insurance), such as making it impossible for certain groups of individuals to access services at a reasonable price. Note that the application of this principle as a core, fundamental principle should be limited to the design of the system, not to its effects. In other words, a system that embeds bias and discrimination "by design" should be considered unlawful in the European Union: a system that is not designed to be discriminatory, but ends up unlawfully discriminating, should trigger liability on the side of the developer, vendor, or distributor. The distinction between the design and the effects of a given AI system is essential in order not to stifle incentives to innovate: the core, inalienable rights to be respected by AI in this first layer are related to the teleological, subjective element of intent, reflected in the design of systems meant to harm or discriminate.
- *Protection of agency, freedom and the democratic process.* Here again, the design of AI systems that have as a final objective, or as an inevitable or easily foreseeable collateral impact, the disruption of individual freedom and self-determination, or the formation of political opinions, should be considered as incompatible with the EU approach to AI. As clarified above, this does not mean that social media like Facebook or search engines like Google should be outlawed since they can be used by third parties to pollute the political debate or generate disinformation (as in the case of Cambridge Analytica, or in the daily practice of "black hat SEO" for search engine optimization). Rather, the interpretation of this principle would point in the direction of introducing a requirement of

[49] Recital 71 of the GDPR states: "In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect."

"enhanced diligence" (compared to the diligence of the ordinary individual or *pater familias*) in managing AI-enabled systems such as online platforms and social media.[50] In short, designing a system to cause harm to individual self-determination or freedom, or even to censor political opinions *a priori* or generate disinformation, should be considered as utterly incompatible with the EU legal system. A different form of protection, based on liability, should be introduced for those systems that, despite not being designed to cause harm, end up infringing these core principles in their concrete implementation, and/or in interacting with other algorithms.

- ***Level 2. Good practices in AI development (Responsible AI).*** This second set would include some of the principles that have been agreed upon by AI developers in documents such as the Asilomar principles. Some of these principles may not be universally applicable, but necessary or desirable in specific applications. For example, the requirement of "human control" over AI (the first of the Asilomar principles), should be included here and explained to developers, vendors and distributors and in the deployment of AI. The extension of the transparency and explainability principles introduced by the GDPR beyond the remit of data protection would also be naturally included in this set of principles. As a preliminary list, level 2 should include the following principles:
    - o *Principle of complementarity with humans ("human-centric AI")*. This principle has been defined in many ways over the past few decades: a "human in the loop" principle would require that a human is always involved in the chain of command that leads to the final output of the AI system; a "human in control" would denote the possibility for human beings to intervene in a timely manner to correct and steer the decision-making process and output of the AI system; and finally, the requirement that there be a "responsible human being" for every act of the AI system is aimed at providing end users with redress in case AI systems cause damage. Of these three possible interpretations, the latter seems to be preferable, although in some specific sectors the need to have a human being in the loop, or in control, may well emerge. For example, in medical diagnostics the need to have a doctor in the loop is essential;

[50] The Asilomar Principles are quite specific on this point, citing the threats of an AI arms race and of the recursive self-improvement of AI, as well as the need for "caution" around "upper limits on future AI capabilities". The Partnership similarly asserts the importance of AI operating "within secure constraints". The IEEE document meanwhile cites the need to "avoid misuse", while the Montréal Declaration argues that those developing AI "should assume their responsibility by working against the risks arising from their technological innovations", echoed by the EGE's similar need for responsibility.

a similar argument can be made for the use of AI-enabled algorithms in court, or by police forces. In self-driving cars, the human in the loop makes very little sense even when it is provided for: not surprisingly, some companies still believe that a lower degree of autonomy, with humans still needed at the steering wheel, is the best approach for the near future. In B2B (business to business) settings, the high level of automation and the amount of data processing may make it difficult to involve a human being in all steps of the supply chain: requiring that this condition be fulfilled may stifle innovation in these B2B settings, thereby frustrating their actual purpose. That said, in most circumstances it should be possible to use AI to augment, rather than replace, human intelligence: this approach to AI should be encouraged to the extent possible, in particular as a "level-3" type of principle (see below). Determining the degree of human involvement in the loop could form part of a preliminary risk assessment, which could be integrated into the overall assessment of liability.

- *Responsible governance: monitoring, control and feedback.* Adequate governance and safeguards in the development of AI systems should be encouraged. This issue is extremely important for Europe, where *ex-ante* controls and enhanced care in developing new technologies are justified also by the lack of a strong litigation culture, which typically constitutes a deterrent from exercising insufficient care in the US in fields such as product liability. Principles of responsible AI for Europe should require (as a non-mandatory element) the implementation of a step-by-step approach to AI development, which checks against undesirable biases in the initial dataset, the curation or cleaning of data, the design of the algorithm, as well as the control of the output, outcomes and impacts of the AI system. This is similar to the so-called "from farm to fork" approach used in agriculture, which aims at providing control points at all stages of the AI development and implementation process.[51] This also helps in tracing responsibility, and collecting feedback, which allows for continuous improvement of the system as well as the responsiveness to user requests and complaints, and making available mitigation measures and remedies.[52]

[51] See https://ec.europa.eu/food/sites/food/files/safety/docs/fs_infograph_from-farm-to-fork_en.pdf

[52] In some cases, this is indeed (or will soon be) mandated by law, for example, under the proposed P2B Regulation (see https://ec.europa.eu/digital-single-market/en/news/

  - *Principle of AI transparency* and *explainability*. AI systems should be transparent when it comes to the data used, the fact that a user interface is not human (so-called counter-CAPTCHA), as well as to the identity of those that trained the system, and basic information on how the system was trained. This, however, does not necessarily extend to all aspects of an AI system, including the intellectual property behind the algorithm, which could well be proprietary.[53] Similarly, AI must be generally explainable, even if the ultimate level of explainability required depends on the expected end users, as well as on the specific use case. For example, in most cases, knowing the identity of the engineer that trained the system would not be a very actionable insight for consumers. While the GDPR introduces a right to an explanation on the side of data subjects, limited to "the existence of automated decision-making, including profiling" and "the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject", this does not automatically imply that all AI systems should be designed as explainable. B2B systems, industrial AI systems, even recommendation systems do not necessarily have to be fully explainable to their users, provided that they do not process personal data without explicit user consent, and that an adequate, well explained liability regime is in place in case they cause harm. In this respect, the GDPR actually creates an island of "level 1" principles within an otherwise "level 2" setting.

- ***Level 3. Principles of sustainable AI (Sustainable AI).*** This set of principles marks the difference between AI that is simply "admissible" in the EU space and AI that is fully aligned with EU policy goals. This space would be mostly occupied by principles of benevolence and alignment with the Sustainable Development Goals, as embedded in the EU 2030 Agenda. They include, as a preliminary list, the following:
  - *"Do good" or benevolence principle.* There should be no mandatory requirement for AI to be designed to foster the global good: as with all products and services, and with the due caveats, the "invisible hand" applies, and the selfish pursuit of profit can provide benefits to society unless proven to do otherwise. This is why the benevolence principle should not be included in the core level 1 principles in the future Ethics Guidelines. This would otherwise rule out the overwhelming majority

regulation-promoting-fairness-and-transparency-business-users-online-intermediation-services).

[53] The EU could then decide to procure only open source AI, but this would then become a level 3 requirement.

of the AI systems used today, which, as explained in section 1 above, are embedded in modular information systems, and not necessarily as a contribution to human development and prosperity. That said, level 3 principles could include a requirement of benevolence, i.e. of AI design to effectively address key societal challenges. This is a general requirement compared to the principles spelled out below, which refer to individual sustainable development goals.[54]

- *Limited or zero carbon footprint (SDGs 7 and 13).* As advocated by some scholars and experts, AI can provide a significant contribution to climate change mitigation, both through the design of specific AI applications that can improve key carbon-emitting technologies; and through the design of systems that have a limited or even zero carbon footprint. The use of renewables to support the energy-hungry activities of data centres should be promoted by EU institutions, and this in turn will provide companies with adequate incentives to develop sustainable hardware and data storage. Benchmarks expressed in the form of data centre power usage effectiveness (the ratio of total power required to run an entire facility versus the direct power involved in computing and storage) could be useful in this respect, coupled with the already-existing European Code of Conduct for Data Centre Energy Efficiency.[55]
- *Inclusive growth, full and productive employment, and decent work for all (SDG 8)*. While AI is generally credited with positive impacts on growth, the jury is out as to whether such growth will be inclusive. As a matter of fact, several authors have flagged the possible problem of market concentration and a future "AI divide", if technology will not be accessible and affordable to everyone. While non-inclusive AI should not, generally speaking, be illegal, the EU may want to encourage the development and implementation of AI products that promote inclusion. Even more controversial would be the development of AI that is compatible with full and productive employment, since it does not lead to the replacement of humans, but rather to their productivity and enhancement in the workplace. Similarly, AI could promote the decency of the jobs to be performed, by liberating individuals from repetitive tasks and fuelling their creativity. But job decency also depends on

[54] A relevant example is the use of AI in advertising, whose purpose is not always aligned with the beneficence principle (although advertising also has an informational component and can have an awareness-raising impact). Restricting the use of AI for marketing purposes based on the "Do Good" principle would have an unintentional constraining effect on AI: advertising, of course, would still have to adhere to Level 1 principles, which include compliance with the GDPR.

[55] See https://ec.europa.eu/jrc/en/energy-efficiency/code-conduct/datacentres.

stability: AI has led to the transformation of many stable jobs into temporary jobs, leading to the emergence of what some scholars call the new "precariat". AI designed with full and decent work in mind could be encouraged and endorsed by the EU in light of its 2030 strategy. In more concrete terms, the following presumptions appear meaningful: fully explainable and transparent AI can be presumed to be more inclusive than less transparent and explainable AI And human-augmenting, rather than human-replacing AI is certainly more compatible with the goal of promoting full and decent employment.

- *Quality education (SDG 4)*. AI can substantially improve access to education and can be designed to this end. For example, breakthroughs in natural language processing and translation, coupled with enhanced connectivity and conversational bots can reduce the cost of accessing top-quality education for all, regardless of the geographical location. It can also improve the accessibility of education for all, including for persons with disabilities. Personalised learning and automated grading will make online education much more compelling and empowering than it is today. The EU could encourage the development of AI systems that help European and global citizens advance towards promoting access to high quality education for all.
- *Promotion of women empowerment (SDG 5*). Outright discrimination based on gender is illegal in the European Union and would thus be a violation of level 1 principles. But AI can unintentionally reinforce gender inequality,[56] thereby making it important to ensure that AI systems are explicitly tested to empower women and control for possible gender biases.[57]
- *Industry, innovation and infrastructure (SDG 9).* AI systems can lay the foundations for further innovation, especially through open data and open IP arrangements. For example, Cockburn et al. (2017) argue that "policies which encourage transparency and sharing of core datasets across both public and private actors can stimulate a higher level of innovation-oriented competition, and a higher level of research productivity going forward".

[56] See Bettina Buchel, "AI could reinforce gender inequality", blog post, at https://www.weforum.org/agenda/2018/03/artificial-intelligence-could-reinforce-our-gender-equality-issues.

[57] See also Katica Roy, "The Economic Case for Using AI to Close the Gender Equity Gap", blog post, July 2018, at https://medium.com/inside-the-salesforce-ecosystem/the-economic-case-for-using-ai-to-close-the-gender-equity-gap-9ba6ac2d4eb6

4.2.2 *Second section: Examples and definition of problematic use cases and "no go's"*

The second part of the Ethics Guidelines could include a list of use cases that are considered to be representative, borderline or problematic by the European Commission. Such use cases could be divided into two sections:

- *A section with "prohibited" use cases.* This would include examples of AI that are incompatible with the fundamental principles listed in the Ethics Guidelines. A relatively well-known one is that of Lethal Autonomous Weapons (LAWs), which have been also the subject of a recent resolution of the European Parliament aiming to ban their uncontrolled use.[58] Their incompatibility with "Level 1 principles" would be a good basis for declaring them unlawful with the EU legal system, unless human accountability is foreseen. The claim that autonomous weapons should remain under human control is almost an oxymoron. A more balanced and credible argument holds that LAWs remain under human responsibility, and thus humans remain accountable for whatever autonomous weapons do. Another example of a "no go" could be the delegation of certain "life or death" decisions to algorithms, without human involvement and control: this could be exemplified by a kidney exchange run exclusively by an algorithm, without human intervention before a decision is made.
- *A section with "problematic" or "borderline" use cases.* As clarified above, the bulk of expected uses of AI poses no outstanding ethical problems. It would indeed be important to maintain a living collection of use cases that are considered to be "problematic" from the standpoint of the European Commission. These use cases could usefully be crowdsourced thanks to the AI Alliance, and then analysed by the European Commission before being included in the living section of the Guidelines. Examples of potentially problematic use cases include the following:[59]
  - *Predictive policing.* Using software such as PredPol requires enhanced attention in collecting, curating and using data and avoiding the amplification of bias at all levels of the process.[60] Explaining how bias can creep in while using predictive policing could help clarify the boundaries of their use in police stations, as well as in regulatory

[58] See *supra* note 47.

[59] A good example of a repository is the AI Now Toolkit for Algorithmic Accountability, which contains a list of existing algorithms being used and distributed in various fields, and for various use cases (see https://ainowinstitute.org/aap-toolkit.pdf).

[60] See https://www.predpol.com/

agencies (e.g. for data-based inspections). Likewise, explaining how predictive policing can lead to violations of an individual's privacy and how to adopt mitigating strategies would help clarify possible actions to be adopted in all similar cases of use of advanced AI-powered data analytics to predict future events (e.g. the likelihood of a child being abused).[61]

- *Social credit scores.* The use of personal data from various sources, in particular from social media, to build and implement a system of social credit scoring should be analysed and ruled out as too intrusive and discriminatory based on EU fundamental rights. The explanation of why social credit scoring is incompatible with the EU legal framework may seem a futile exercise, but it can provide useful guidance in all those cases that already widespread in Europe, in which data from social media are being used to discriminate between end users, for example in insurance services. To what extent should it be possible to use available public information to this end? How should trade-offs be solved when algorithmic accuracy is enhanced by the intrusive use of data from social media? Would it be possible in Europe to entice users into more attractive products or services in exchange for (explicit consent to) access to their personal data, as in emerging markets such as those for direct-to-consumer genetic tests?
- *Facial and body recognition.* It goes without saying that the use of facial and/or body recognition can increase the effectiveness of police enforcement. Recent cases have shown that AI can spot criminals among thousands of people, e.g. in a stadium. Advanced image recognition and rendering techniques can also lead to identifying criminals starting from very blurred images. In a time of constant risk of terrorist attacks, massive use of facial recognition is too attractive to be discarded all at once. That said, what are the limits to the use of this technique in the

[61] Automating Inequality includes a discussion of the Allegheny Family Screening Tool (AFST), a predictive risk model deployed by the County Office of Children, Youth, and Families to forecast child abuse and neglect. While the AFST is only one step in a process that includes human decision-makers, Virginia Eubanks argues that it makes workers in the agency question their own judgment and "is already subtly changing how some intake screeners do their jobs". Moreover, the system can override its human co-workers to automatically trigger investigations into reports. The model has inherent flaws: it only contains information about families who use public services, making it more effective at targeting poor residents. Such discriminatory effects cause harm in other human rights areas, such as education, housing, family and work. See Virginia Eubanks, *Automating Inequality,* London: St. Martin's Press, 2018 and also Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, NY: New York University Press, 2018.

public and private sectors? Could a private corporation use the same facial recognition technique used to spot criminals to enable new services in social media, such as matching people with places and advertisers? Could facial recognition be used in combination with other datasets to determine a person's likelihood to repay a debt, and accordingly reach a decision on a user's creditworthiness?

- *Content filtering.* Increasingly, regulators are relying on online intermediaries to enforce legal rules, including those on hate speech, disinformation, and copyright. Algorithmic take-down of online content can potentially undermine freedom of expression. In the absence of a strong legal framework on this fundamental right, however, online intermediaries inevitably err on the side of "Type 1 errors" or "false positives" (i.e. if there is any doubt, they take down content even if it does not infringe the law, in order to avoid the risk of incurring liability). Digging deeper into this use case of AI would provide more certainty to online intermediaries in many similar settings, when they have to ensure algorithmic compliance with the law by adopting best-practice behaviour.
- *Conversational bots.* Using conversational bots can increase the efficiency of specific services and even improve user experience in most cases. At the same time, however, there is a risk of discrimination and deteriorating quality of service, which should be mitigated through specific actions. Describing this use case, clarifying which actions are lawful and recommended, would improve legal certainty for a large number of future applications. It would also provide the European Commission with the opportunity to clarify the circumstances in which end users should be given a clear explanation of the non-human nature of the interface and the possibility to interact with a human being rather than with a bot.

4.2.3 Third section: Guidance on mitigating strategies (good practice)

Rather than simply flagging problematic cases, the Ethics Guidelines could also point at possible arrangements that would significantly mitigate the risks posed by a specific use of AI. This section, too, could start with a collection of good practices, and be then open to ongoing submissions through the AI Alliance, of examples of remedies that have successfully mitigated the problem at hand.

In the case of personal data, this has already occurred in many ways, mostly through private regulation, certification and the adoption of Privacy Impact Assessment practices. In addition, there is reason to believe that the **first strongholds against bias creeping into corporate algorithms come with embedding good practice into companies' daily risk management activities**.

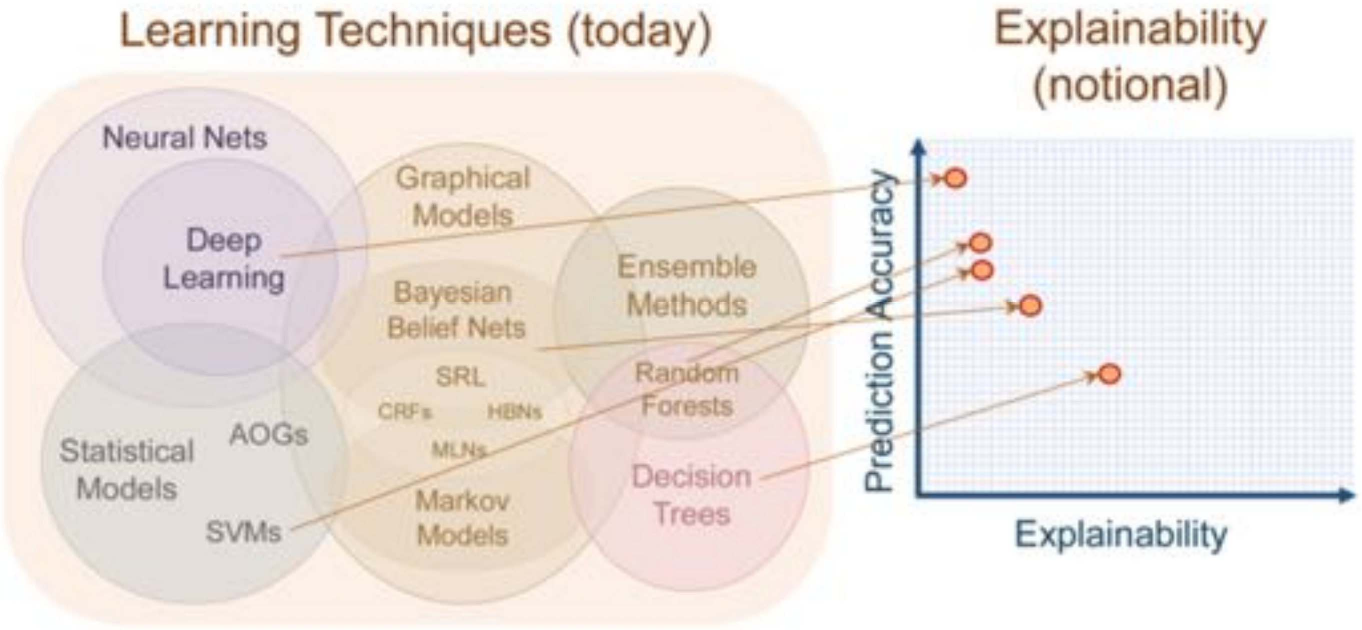
In this respect, this section of the guidelines could incorporate a tripartite analysis of companies' lines of defence: i) operational management; ii) risk management and compliance functions; and iii) internal audits. Consultancy firms like Accenture have already developed a full "AI fairness tool", based on the work of data scientists on quantitative fairness, which can be incorporated into this framework, leading to a more effective and agile governance of risk in the corporate environment. Similarly, IBM developed tools such as AI Fairness 360 and the Everyday Ethics Guide for AI, which provide concrete solutions for developers and users. While these systems do not necessarily solve all problems when it comes to de-biasing and value alignment, their application to AI design and distribution could mitigate the risks associated with the infringement of Level 1 and Level 2 principles.

Moreover, guidance could be offered to developers, vendors and SMEs as to **when techniques such as unsupervised deep learning and reinforcement learning are appropriate**, and what are the consequences in terms of explainability of AI. Figure 9 sketches the trade-off that companies face between explainability and accuracy of algorithms. In this field, private-sector guidance is already advancing rapidly and is expected to further improve and expand in the near future. For example, Google's Tensorflow recently released a "what-if" tool to visually inspect machine-learning models, within the People + AI Research initiative (PAIR).[62] These systems are able to show the behaviour of the model (as black box), and should gradually move towards a full explanation of how the system reaches decisions, and even more importantly, how the system reached a given decision, for which an end user awaits explanation.

Finally, guidance could also be offered with respect to making algorithms GDPR-proof by using specific cryptographic techniques. While most of these arrangements may already be embedded in pre-trained algorithms sold by IT companies on the market, developers should constantly engage in discovering and testing new techniques to promote the protection of users' personal data in AI systems. Current, widely tested techniques include *zero-knowledge* proof systems, in which a prover convinces a verifier via an interactive protocol that some statement is true, i.e. a given word x is in some given language L; and *homomorphic encryption*, which allows computation on ciphertexts, generating an encrypted result which, when decrypted, matches the result of the operations as if they had been performed on the plain text.

[62] https://pair-code.github.io/what-if-tool/

Figure 9. The explainability-accuracy trade-off



Source: Zelros AI, "A Brief History of Machine Learning Models Explainability", Medium, 24 September 2018 (https://medium.com/@Zelros/a-brief-history-of-machine-learning-models-explainability-f1c3301be9dc).

4.2.4 *Fourth section: What to do in case of…*

The fourth and last section of the Ethics Guidelines could include guidance for the end users, integrated with existing projects (e.g. AlgoAware) and with online tools to educate end users about their rights and what to expect when dealing with AI-enabled systems. This section, too, could be usefully fed by the contributions from end users and other stakeholders (e.g. SMEs, developers) thanks to the AI Alliance and future AI Observatory. A useful way to provide guidance to end users would be to include interactive materials on how to react in specific circumstances, e.g. in case a given AI system has not given them enough information or has come to a questionable decision without leaving the possibility to obtain clarification.

4.3 An analysis of the current Draft Ethics Guidelines

In a working document released on 18 December 2018, the European Commission presented the first draft of the AI Ethics Guidelines prepared by the AI HLEG. The working document does not represent the final position of the HLEG, and it was clarified that "a number of themes still need to be elaborated in more detail". The European Commission also explained that the final version of the Ethics Guidelines will feature a mechanism for stakeholders to voluntarily endorse the Guidelines; and that only at a later stage will it examine whether to formally adopt the Guidelines in an official document, and/or whether additional measures may be needed to deal with the ethical challenges of AI.

That said, the published document already presents a relatively clear picture of where the HLEG is directing its efforts in advising the European Commission. **The HLEG aims at maximising the benefits of AI while minimising the risks; and chooses "trustworthy AI" as the goal for Europe's approach and policy**. Trustworthy AI has two components: i) its development, deployment and use should respect fundamental rights and applicable regulation, as well as core principles and values, ensuring an "ethical purpose", and ii) it should be technically robust and reliable.

The CEPS Task Force on AI believes that the future Ethics Guidelines could represent a very welcome development at the EU level and could also structure its work and discussion with a view to contributing to this ambitious and highly deserving endeavour. Accordingly, we offer some observations in the following pages on the first draft of the Guidelines published on 18 December 2018.

4.3.1 Values and principles: The need for a hierarchy

The Draft Guidelines observe that AI should be **human-centric**, and thus be "developed, deployed and used with an 'ethical purpose', grounded in, and reflective of, fundamental rights, societal values and the ethical principles of Beneficence (do good), Non-Maleficence (do no harm), Autonomy of humans, Justice, and Explicability".[63] AI developers should also rely on fundamental rights, ethical principles and values to prospectively evaluate possible effects of AI on human beings and the common good; and pay particular attention to situations involving more vulnerable groups such as children, persons with disabilities or minorities, or to situations with asymmetries of power or information, such as between employers and employees, or businesses and consumers. They should also be aware of the fact that AI can have a negative impact and remain vigilant to avoid it.

Compared to the ethical guidelines proposed in section 4.1 above, **the current list lacks hierarchy between principles and values that should always be complied with as they constitute the core constitutional principles of the EU and those that correspond to good practices in AI development or are aligned with EU medium-term policy goals**. This differentiation would add considerable value to the Guidelines, which otherwise would end up re-proposing the same structure and rather vague list of principles already endorsed by the EGE group, and also by many other governmental and non-governmental documents over the past few years. Developers, vendors, distributors and users would not be able to learn from the list of values and principles which forms of AI are to be considered lawful, and which are not. In particular, including the beneficence principle among the core ones, as in bioethics, appears to be disproportionate, and simply observing that developers should be aware of the fact that AI can also have a negative impact also seems a very difficult principle to verify and enforce.[64] The Draft Guidelines end up explaining that "AI systems should be designed and developed to improve individual and collective well-being". But whether this should be a mandatory intent of AI designers and developers is highly questionable, and any alternative explanation based on the pursuit of profit would fall short of complying with the "ethical purpose" requirement.[65]

[63] See the Draft Guidelines, Chapter 4 (https://ec.europa.eu/futurium/en/node/6044).

[64] The Draft Guidelines explicitly refer to the Council of Europe "Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine" (the Oviedo Convention).

[65] As a matter of fact, the Draft Ethics Guidelines propose that AI development, deployment and use should serve an "ethical purpose", while defining ethics as a field of study that centres

Also, **the principle of non-maleficence, which this report also listed among the "level 1" principles, is currently defined in a way that is hardly actionable for developers, distributors, users and policy-makers**. More specifically, the Draft Guidelines observe that "AI systems should not harm human beings", and that "At the very least, AI systems should not be designed in a way that enhances existing harms or creates new harms for individuals". If further clarifies that harms can be physical, psychological, financial or social, and can also constitute the unlawful treatment of personal data. However, more clarity would be needed for all those borderline cases in which an AI system ends up in a "life or death" situation or even more likely in a "death or death" type of dilemma (as in the trolley problem). Would this constitute a new harm, a greater harm or an old harm? Would the use of a conversational bot that changes children's way of interacting with machines and even their accent when they talk violate this principle, by creating psychological harm (Yi Cheng et al. 2018)? What will be the boundaries between harm that is pre-existing, and new harm? And between harm that is "lawful" and harm that infringes one's rights?

The same principle (do no harm) also contains environmental-friendliness as a principle that echoes that of sustainability, which this report included as a main "level 3" principle. However, there are two potential issues with the choice made in the Draft Ethics Guidelines. First, **sustainability not only has an environmental dimension, but it also features important social and economic aspects** that are not fully covered in the Guidelines. Second, **including environmental sustainability under the "do no harm" or non-maleficence principle does not necessarily mean enforcing these principles in practice**. More specifically, stating that AI systems should comply with environmental sustainability could either result in the introduction of a new minimum standard (i.e. non-compliant AI systems would be illegal once the principle is translated into bunding legislation); or in the simple description of what constitutes "ethical" AI, with no real enforcement solution attached. Based on the current text of the draft Guidelines, the latter solution seems most likely.

In the section dedicated to the **preservation of human agency, or self-determination**, the Draft Guidelines mention that "if one is a consumer or user of an AI system this entails a right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal". This is said to include a right to "individually and collectively decide on how AI systems operate in a working environment", and relatedly, "provisions designed to ensure that

on questions such as: "What is a good action?", "What is right?", and in some instances "What is the good life?".

anyone using AI as part of his/her employment enjoys protection for maintaining their own decision-making capabilities and is not constrained by the use of an AI system".[66] Here, too, the boundaries and enforceability of these provisions should be clarified.

On the one hand, the right to opt out from a system that uses AI as support to decision-making can be very far-reaching. For example, while the right to refuse direct interaction with AI is more commonly invoked, the right to refuse indirect interaction could mean that border control officers, tax enforcement agencies or medical doctors could be asked to refrain from using specific software when interacting with a specific customer. It may also imply that employers have to comply with specific standards when setting up human-machine cooperation in the workplace, if they do not want to infringe occupational safety and health legislation. These are just examples, but the bottom line again is that **these provisions read like "all or nothing": either they are fully enforced, leading to massive over-regulation (standards, certification, safety at the workplace, obligation to train personnel to provide services with and without AI support) or they are not made enforceable and they might simply remain "wishful thinking", like many other declarations of principles and values**.

Similar issues emerge with the principles of **fairness** and **justice**. The draft text implies that the development, use and regulation of AI systems must be "fair", and this entails that the "positives and negatives resulting from AI should be evenly distributed". While the remainder of the text is highly aligned with the analysis proposed in this report (i.e., justice implies providing users with effective redress and being held to high standards of accountability), **the idea that all AI systems should be gauged against fairness standards should be accompanied by extensive explanations on key aspects of AI development and commercialization, such as what is acceptable bias, when does it become unacceptable (see above, section 1) and ultimately, what is a fair outcome depending on the specific use case?**

This section is followed by a **very useful section (no. 5) on critical concerns raised by AI**, which – despite lack of consensus among experts on certain issues – are meaningfully explained and presented. That said, the CEPS Task Force proposes that the HLEG looks at our section 4.2.2 above for additional insights on issues that can be representative of common issues that are raised by use of AI systems in various settings.

[66] Footnote 13 of the Draft Ethical Guidelines and associated text.

4.3.2 *Implementing trustworthy AI: Giving more "teeth" to the Guidelines*

The Draft Guidelines contain a section on implementing trustworthy AI, which in principle should address some of other concerns raised in the previous section about the lack of enforceability of the core values and principles identified by the HLEG. The Group itself observes that "achieving trustworthy AI means that the general and abstract principles need to be mapped into concrete requirements for AI systems and applications". The first step chosen by the HLEG is the mapping of the values and principles into ten "requirements", presented as non-exhaustive and in alphabetical order.[67]

1. **Accountability** mechanisms are briefly presented, mostly in the form of legal liability (e.g. being responsible to compensate for damages or to offer apologies). There is no mention of the relationship between accountability and liability, and no discussion of the type of legal liability that is required for AI systems. The treatment of compensation of moral damages is absent and is briefly replaced by a statement according to which "in a case of discrimination, however, an explanation and apology might be at least as important".
2. **Data governance** refers to good practice in the collection, division, curation and integrity of data used in AI.
3. **Design for all** implies, in the interpretation of the HLEG, the "accessibility and usability of technologies by anyone at any place and at any time, ensuring their inclusion in any living context, thus enabling equitable access and active participation of potentially all people in existing and emerging computer-mediated human activities". However, AI systems are often designed for specific users, in particular professional users, e.g. AI used in support of medical doctors for better diagnoses or used in industrial B2B settings. **Requiring AI to be always designed for all users appears to be disproportionate at best, and likely to impose unnecessary costs to all developers of professional AI systems, or even AI design to augment humans in specific professions**.
4. **Governance of AI autonomy (human oversight)**. This requirement broadly implies that the greater the degree of autonomy left to the AI system, the stricter should be the governance built around it, and the requirement that users, especially in a work or decision-making environment, are "allowed to deviate from a path or decision chosen or recommended by the AI system". The relationship between this

[67] Some members of the CEPS Task Force are of the opinion that the list of requirements is at once too long, and redundant, given the many overlaps between principles such as transparency and accountability, as well as robustness and safety.

requirement and the issue of human-centricity is not clear and raises doubts as to whether ethical AI could operate in the absence of a human in the loop (but still in the presence of a human responsible).

5. **The requirement of non-discrimination** echoes a concern that is widely shared among experts and should possibly be coupled with an **explanation of what constitutes acceptable discrimination and what forms of discrimination are unacceptable**. For example, in economic theory, discrimination is normally considered to be welfare-enhancing, as it can lead to enhanced market access and a reduction of the deadweight loss to society associated with -imperfectly-competitive markets (Armstrong 2005). Is AI that achieves perfect price discrimination unethical? For example, would an auction system that leads bidders to offer a price as close as possible to their willingness and ability to pay be considered unethical? What about an e-commerce platform that engages in personalized pricing based on observable information (e.g. how long has the user been surfing)?[68] Clarifying those aspects is exactly what the Draft Ethics Guidelines should do to increase the level of certainty and trust in the way AI interacts with society and the economy.
6. **Respect for (and Enhancement of) Human Autonomy.** This requirement, deeply linked with no. 4 above, is illustrated with respect to so-called "extreme" personalisation approaches, such as "recommender systems, search engines, navigation systems, virtual coaches and personal assistants"; these systems should, according to the Draft Guidelines, prioritise the overall wellbeing of the user. In this case as well, however, **the exact definition of user well-being and the cases in which the system excessively nudges the user are not clear**.
7. **Respect for privacy** is very briefly described as a requirement for organisations to be "mindful of how data is used and might impact users and ensure full compliance with the GDPR as well as other applicable regulation dealing with privacy and data protection". This is already a legal obligation, which makes it **even more important to couple it with an indication of best practices and a description of specific use cases that could improve the awareness of AI designers, developers, vendors and distributors when it comes to specific forms of privacy protection, especially "by design" solutions**.
8. **Robustness** is related mostly to security, resilience but also to transparency and replicability of results. The latter is the most

[68] On the pros and cons of personalized pricing, see also the proceedings of the meeting of the OECD Competition Law and Policy Committee (http://www.oecd.org/daf/competition/personalised-pricing-in-the-digital-era.htm).

interesting and controversial, since the HLEG argues, in the current draft, that trustworthy AI requires that "the accuracy of results can be confirmed and reproduced by independent evaluation". This obviously raises the issue of so-called "black boxes", which can reach high levels of accuracy through deep neural networks, but at the price of explainability. The draft Guidelines seem to imply that trustworthy AI can be reproduced and that "lack of reproducibility can lead to unintended discrimination in AI decisions". Again, **it is not clear how far-reaching this requirement is**. Should, for example, AI systems used in B2B settings lead to fully reproducible results as a requirement for trustworthy AI?

9. **Safety** refers to the protection of human integrity and seems to hint at similar provisions to the ones already in force under the EU on Products Liability. However, given that trade-offs may emerge between, i.a. explainability and effectiveness of AI, it would be essential to provide guidance on how to strike the right balance in given use cases. As such, any assessment of AI safety should be sufficiently nuanced to reflect this issue.[69]
10. **Transparency** introduces the principle of explainability as an element of trustworthiness. In particular, in case the AI system "uses human data" or "affects human beings" or have "other morally significant impacts", this requirement entails that the system be explicit and open about choices and decisions concerning data sources, development processes and stakeholders. Here too, **the main message of the HLEG is that only explainable AI is trustworthy AI: but it remains unclear what happens to non-explainable AI under EU law. Is it lawful and unethical? Or unethical, but still lawful?**

4.3.3 *Guidance for ensuring and assessing trustworthy AI*

The remainder of the AI HLEG's Draft Ethics Guidelines is dedicated to illustrating good practice to ensure trustworthy AI; as well as providing key guidance on assessing AI trustworthiness. The main content of these sections focuses on the need to incorporate trustworthiness early on in the design of AI and to consider both technical and non-technical methods to ensure the implementation of the aforementioned ten requirements in AI systems. The ten

[69] Andrew Tutt (2017) notes that "…when and why machine-learning algorithms fail is difficult to predict and explain because what they do is probabilistic and emergent by design". He also notes AI's similarity with drugs, in that "…the precise mechanisms by which they produce their benefits and harms are not well understood…" As such, he proposes a Federal Regulatory Agency that oversees AI in much the same way as the US Food and Drug Administration oversees pharmaceuticals.

requirements should also be kept in mind when training teams that will work on the system, the system itself, the testing environment and the potential applications of the system. AI designers and developers are also asked to provide, in a clear and proactive manner, information to stakeholders (customers, employees, etc.) about the AI system's capabilities and limitations, including trade-offs, in order to allow them to set realistic expectations.

AI designers and developers are also advised to "adopt an assessment list for Trustworthy AI" and to adapt it to the specific use case in which the system is being used. Such a list may not be exhaustive, and in any case, trustworthiness should be subject to a cycle of assessment, monitoring, learning and updating. Such a governance cycle should be coupled with a "human-centric approach to Artificial Intelligence, which will enable Europe to become a globally leading innovator in ethical, secure and cutting-edge AI". It should also strive to facilitate and enable "Trustworthy AI made in Europe", which will enhance the well-being of European citizens.

The Guidelines end with a relatively long list of assessment questions on trustworthy AI. These are based on the "ten requirements", which effectively take the lion's share of the Guidelines. This is why it is of the utmost importance that **an in-depth discussion takes place on how to deepen guidance on these requirements, as well as on the "elephant in the room": What happens if a given AI system fails to meet the specified requirements?**

4.4 "How" to promote the Ethics Guidelines?

The four-section structure we propose for the future Ethics Guidelines in section 4.2 above differs from the current structure proposed by the HLEG in the first draft published in December 2018. In what follows, we argue that the four-section structure we propose would significantly improve on the current draft, as it would mark a significant departure from existing documents, including AI principles; and it would also make the guidelines future-proof, due to the following characteristics:

- **Scalability**. The structure of the proposed Ethics Guidelines as integrated in an online, interactive tool would give the European Commission the possibility to add and update content under the different sections, and crowdsource content from the AI Alliance, maintaining it as a constant source of information and guidance.
- **Flexibility.** The three-level structure of the principles allows for clarification of the "red zones" or "no-go's", which correspond to the Level 1 area, plus the general or sectoral legislation that directly prohibits specific conduct (e.g. the GDPR for cases of processing of personal data); and at the same time allows for the promotion of more

responsible AI practices, to put the first in line with generally acknowledged principles of responsible AI, and then in line with the EU 2030 Agenda. Moreover, the guidelines are also very flexible since they allow for the reclassification of specific principles across levels, following the entry into force of new legislation.

- **Modularity.** The modular structure, with four sections and a three-tiered list of principles, allows for a gradual upgrade and innovation without disrupting legal and regulatory certainty.

That said, even if the Commission will not decide to amend the structure of the Guidelines, it should at least introduce a hierarchy across the ten requirements and provide additional information on how to respond in case trade-offs arise between the principles on which they are based. Apart from the design and content of the Guidelines, a very important separate issue is **ensuring that they have an effective and positive impact on the AI market in Europe**. In this respect, there are several options that could be followed by the European Commission. These include the following: i) making specific principles mandatory, and subject to *per se* rules in adjudication (i.e. there would be no need to prove the effect, but only the existence of the practice); ii) establishing rebuttable presumptions of unlawfulness in case of failure to comply with specific principles; iii) reserving specific procurement markets for AI systems that comply with Level 2 and/or Level 3 principles; iv) requiring specific Algorithmic Impact Assessments in specific sectors, based on the use cases thought to be problematic; v) creating an "EU AI mark" that certifies the compliance with some or all the principles contained in the Guidelines; vi) establishing a co-regulatory regime that delegates to industry the development of means to achieve compliance with the principles over time; or vii) leave the issue to self-regulation, encouraging industry players to align with EU values through soft law and moral suasion.

Another important means of promoting the guidelines would be to **propose the adoption of some of its principles in global for a, such as the OECD and the G20**. The scalable and modular structure proposed above could also provide a very flexible scheme for non-EU countries wishing to align their AI strategies and rules with the EU's approach. Access to the EU market would be guaranteed via compliance with level 1 principles, but full alignment with EU values and goals (and possibly, access to certain procurement markets) would occur only after full compliance with level 3 principles was demonstrated.

In reality, the European Commission will not be in a position to adopt only one strategy. **A combination of actions will be required to fully achieve the goals stated in the Communication on AI adopted in April, and the Coordinated Plan announced in December**. The most frequently recurring proposals are discussed below, based on the reflections of the Task Force members.

4.4.1 Is AI the "new GDPR"?

One of the most frequently heard statements in the European debate on AI is that Europe could become a leader in this domain if it manages to make its approach to AI mandatory, applied to all entities that want to operate in the European market, regardless of where they are from. This is based on the alleged success of the General Data Protection Regulation, which effectively introduced a novelty to the world of internet policy by presenting a systematic, well-structured legal framework for the protection of personal data in the European Union, inducing many companies around the world to consider measures to comply with its provisions. Since the GDPR only recently entered into force (25 May 2018), it is too early to judge whether its relatively strict provisions will become a global standard, or even whether they will be fully and homogeneously complied with in the member states. As a result, invoking AI as the new GDPR can only convey a general message: that the EU is entertaining the introduction of heightened mandatory standards on AI to establish a minimum level of protection for end users in the AI age, and will leverage the sheer size of its internal market, as well as the current lack of comprehensive regulation in other large countries, as a way to impose its principles as global "golden rules" on AI.

While the "AI as the new GDPR" vision is interesting, it is not necessarily compelling. After all, it is too early to conclude that the GDPR has become a global standard, and it is also too early to assess the impact of GDPR on Europe's competitiveness in the digital sphere. A proportionate approach would be needed, aimed at steering AI towards the common good when needed, but also careful not to overburden innovators and entrepreneurs with procedural requirements and compliance costs. Already providing guidance to AI developers, vendors and distributors on how to ensure that AI is compatible with the GDPR would be a major step forward (Kingston 2017): in this respect, rather than presenting AI as the new GDPR, it would be important to ensure that AI and GDPR go hand in hand in Europe.

4.4.2 Should an "AI seal" and related mandatory certification system be introduced?

One important aspect of the current debate on the EU approach to Artificial Intelligence revolves around the possibility that the Ethics Guidelines would become a benchmark for certification. Such certification could either be binary (i.e. compliant or not compliant) or layered, in line with existing certifications that provide more granular information as regards the degree of alignment of a specific AI system with the Ethics Guidelines. In addition, such certification could be introduced by the European Commission, similar to recent developments in the field of cybersecurity;[70] or developed by the private sector (in the same vein as the EU Data Protection Code of Conduct for Cloud Service Providers[71]) as a way to build user trust in AI, and therefore signal what would otherwise remain as a so-called "credence" quality in AI.[72] Either way, certification could help reduce the existing informational asymmetry between the suppliers and the users of AI, signalling the existence of ethics "by", "in", and "for" design in given products and services. Certification could also be usefully coupled with standards, in particular emerging IEEE (Institute of Electrical and Electronics Engineers) standards on how to align Artificial Intelligence systems with ethical values; and possibly, future ETSI (European Telecommunications Standards Institute) standards on how to align AI systems with EU "level 3" principles.

However, there are disadvantages in imposing standards at a very early stage in the development of a given industry or family of technologies: early standardization can create a straight-jacket effect, forcing technological development into a pre-determined direction, which is then self-reinforced through path-dependency; this, in turn, creates a lock-in effect, which may lead the AI community into sub-optimal standards.[73] Accordingly, **the decision whether to impose early standards on AI, rather than public or private certification, must be approached with extreme care**.

At this stage, **the CEPS Task Force did not find sufficient grounds to suggest the adoption of public certification, or even mandatory standards on AI in Europe**. This finding is based on three main observations. First, the market seems to be generating self-certification frameworks and packages, mostly through large consultancy firms and tech companies, some of which are extremely active in producing step-by-step guidance and solutions for firms that adopt AI-enabled systems. Second, one would expect that the adoption of the EU Ethics Guidelines will provide this blossoming market with an additional benchmark, leading consultants and intermediaries to signal to their customers

[70] https://ec.europa.eu/digital-single-market/en/eu-cybersecurity-certification-framework

[71] https://eucoc.cloud/en/about/about-eu-cloud-coc.html

[72] Michael R. Darby and Edi Karni (1973) have introduced the term "credence good" and added this type of good to Phillip Nelson's (1970) classification of ordinary, search and experience goods. See, i.a., Dulleck et al. (2011) for a comprehensive economic analysis.

[73] The concept of lock-in was originally developed by Arthur (1994), who discussed the outcome of competition among technologies in the presence of increasing returns to adoption, as we would expect the case of AI would feature, thanks to network externalities and the modular nature of AI systems.

the alignment of their product with most if not all of the principles put forward by the HLEG in the Draft Ethics Guidelines. Importantly, relying on intermediaries would also have the advantage of guiding companies adopting AI through a double process: the value alignment of their AI products, and the alignment of their whole organization with ethical principles, which appears necessary especially for data protection purposes. Already companies like SAP, Deutsche Telekom, Google, Telefonica and IBM are signalling to the market their willingness to go beyond existing ethical principles in dealing with AI solutions: for example, some of them have already officially committed to adopt a full value chain approach, by working with vendors, suppliers and other players along the value chain on the condition that they, too, are aligned with the Guidelines. This "viral" nature of Ethical AI is potentially a very promising avenue for spreading good practice across the industry.

Third, it is still too early to anticipate with reasonable certainty how the AI market will develop over time, and in various domains and sectors. One possibility is that IT firms and consultancies will end up supplying most of the AI solutions in the form of pre-trained algorithms, for use by SMEs. Should this be the case, then the burden of value alignment of AI systems would probably be less significant, since these firms would be able to exploit economies of scale in the design of their systems, and it will not be up to SMEs to perform this rather complex set of activities. Similarly, "ready-made" AI products foreseen by the European Commission in the AI-for-demand platform could also usefully include fully compliant AI solutions, and this in turn may further boost the impact of the Ethics Guidelines on the whole AI community.

All in all, **the preferred solution at this stage seems to be the imposition of ethically aligned AI ("level 3" AI) only in very specific contexts, such as the AI-on-demand platform, public procurement of AI solutions for public administrations and the delivery of public services; and the use of AI solutions in research, innovation and investment policy (e.g. in Horizon Europe, in the future InvestEU and in cohesion funds)**. The European Commission could also consider imposing fully ethically aligned AI in specific sectors, such as healthcare, but only after performing a careful impact assessment. The European Commission should also promote the Ethics Guidelines by mobilizing the AI alliance, coordinating with member states and raising the awareness of the end users through examples and use cases that explain the importance of ethically aligned AI. Together with these activities, the European Commission should also monitor the market to ascertain whether the Guidelines are having the expected impact. Only if the Commission realizes that these measures are insufficient to steer the market towards accountable and sustainable uses of AI, should more intrusive measures such as co-regulation be considered.

4.4.3 *Monitored self-regulation, experimentation and private governance: Let ideas flourish*

In a constantly and rapidly changing environment, the reversibility and adaptive nature of policy approaches are of the utmost importance. As explained above, the creation of lock-in and sub-optimally path-dependent policy solutions can create long-lasting, often irremediable problems for an industry as well as related foregone opportunities for society. In the case of AI, the cross-cutting nature of this industry, coupled with the fact that most of the existing applications present no significant ethical dilemmas (even if they still need to comply with EU law), warrants a very cautious approach in deciding what to regulate, and how. As a result, **regulating to impose full alignment with the future Ethics Guidelines appears disproportionate at this stage**, and likely to impose significant administrative burdens and enforcement costs, and thereby placing the whole EU industry at a disadvantage. As already explained, **the "AI as the new GDPR" appears far from compelling to the members of the CEPS Task Force: rather there was general agreement that the GDPR and other broad existing legislation, such as the Network and Information Security (NIS) Directive, should remain the only mandatory, strictly enforced part of AI, in addition, of course, to "level 1" Ethics Principles**. Other policy measures that may be needed, which are not necessarily related to the Ethics Guidelines, are discussed in the next section, particularly concerning liability rules.

Therefore, for purposes of promoting the Ethics Guidelines, the CEPS Task Force supports the so-called "monitored self-regulation" option. The European Commission should first evaluate the services offered to EU citizens today and their likely evolution over time. And only if it emerges that those services are unlikely to comply with the Guidelines, should the Commission consider more policy actions. This does not mean, however, that the EU should not take any policy measures, for example, to clarify the issue of liability for damages caused by AI systems or to encourage data-sharing in specific sectors. These measures, notably, should imply **adequate room for experimentation**, in the form of "regulatory sandboxes" or randomized controlled trials to ensure that AI-enabled solutions that potentially pose ethical concerns prove their potential in terms of user protection before being admitted to the market. **Experimentation should also be given a very prominent role in testing the Guidelines and in particular the assessment checklist it contains**.

5. POLICY CHANGES: REVOLUTION OR EVOLUTION?

The second deliverable of the EU High-Level Expert Group on AI will be dedicated to the formulation of recommendations in terms of policy and investment. This section briefly looks at policy changes needed to promote and implement the European approach to AI, starting from general considerations about the EU Better Regulation agenda (section 5.1); and then digging more in-depth into the issue of reforming the Product Liability Directive and the Machinery Directive (section 5.2); intervening to facilitate data sharing as a horizontal policy (section 5.3); and with sector-specific policies (section 5.4). The section ends with a brief reflection on governance issues, particularly the merit of existing proposals to create a regulatory agency for AI, data or machine learning (section 5.5).

As a preliminary statement, **the CEPS Task Force did not find strong evidence that would favour a massive revision of existing horizontal regulations in fields such as product liability and machinery**. On the contrary, the members identified a need to strengthen the experimental dimension of EU regulation, and reconcile three apparently incompatible stances: the need for precaution when allowing the commercialization of certain new technological solutions; the fact that most AI does not pose ethical or regulatory concerns that are significantly different from those posed by existing IT solutions; and the need to test new solutions that have the potential to positively contribute to societal welfare and sustainable development. Accordingly, the answer to the question in the title of this section appears to point in the direction of an **evolution, rather than a revolution**. It must be recognised, however, that technology is also bringing important changes in the way governments approach regulation, and this will affect the way in which AI policy is likely to evolve over time.

5.1 The Better Regulation agenda and AI

The EU can rely on a very well-structured, sophisticated Better Regulation agenda, which also includes ad hoc tools on measuring impacts on innovation, as well as assessing issues related to the Digital Economy and ICT.[74] The Better Regulation Guidelines provide an almost complete tool box that the European Commission and other EU institutions can safely rely upon in deciding what policy approach to adopt in the face of emerging digital technologies, such as AI. The only areas that seem to be partly absent in the Better Regulation toolbox are behavioural economics (including "nudging") and experimental policy-making. Guidance would be needed for European Commission services to take into account the behavioural reactions of end users, and how to ensure that regulation, when needed, is adaptive, flexible and innovation-friendly.

That said, it would be important that when reviewing existing legislation, and even more importantly when preparing the ex-ante impact assessment of new policies, the European Commission takes into account the following:

- The **problem definition should be increasingly based on foresight and risk analysis**, as well as the result of fact-finding with a large group of stakeholders. This could be, in the future, the role of the AI Alliance and Observatory. The latter could complement the role of the REFIT platform in signalling cases of legislation in need of revision.
- The Commission should remain **open to a new form of "innovation deals"**, in which AI developers challenge existing regulation by showing that they can achieve more significant benefits and higher levels of user protection with alternative business models than the ones on which the original Regulation was based. This could be done in a more effective way if a mission on AI (or a "mission IT", see below section 5.4.4) were launched; and even more if the Commission will strengthen the experimental dimension in the Better Regulation agenda: as a matter of fact, innovation deals can easily lead to the launch of an experimental phase.
- **An experimental phase should be foreseen, in which new co-regulatory solutions and/or new business models are tested in a secure space** such as a sandbox, before they are admitted to the market. The sandbox should be designed in a way that does not entail excessive administrative burdens for the entrepreneurs willing to enter the market with innovative solutions. At the same time, the sandbox should not compromise the safety and fundamental rights of the end users, and as such should not violate "level 1" principles.
- In terms of **methodology**, the Commission should analyse possible changes in the AI-related policy framework by **avoiding exclusive reliance on cost-**

[74] See https://ec.europa.eu/info/sites/info/files/file_import/better-regulation-toolbox-27_en_0.pdf

benefit analysis.[75] **Rather, a multi-criteria analysis would be more appropriate**: regulatory options should lead to a high level of protection of basic and fundamental rights, as defined above under Level 1, as well as additional policy objectives, such as user empowerment, and the 2030 Agenda sustainability goals.

- **Concerns alternative policy options, it is very important that the Commission respects the so-called "Treaty-based principle of proportionality"**, which dictates that any proposed means of intervention should be proportionate to the stated goals, and thus avoid being overly prescriptive or invasive. In particular, the Commission should consider the following alternatives when deciding on possible policies in AI-intensive fields:
    - *Awareness raising campaigns*. These should be designed in order to reach and empower end users, at the same time exploiting behavioural science in order to maximise their effectiveness.
    - *Self-regulatory schemes*. these include codes of conduct for algorithm developers (e.g. the Asilomar principles), corporate privacy/data protection policies, privately managed complaint handling and enforcement mechanisms (e.g. take-down policies, policies on the right to be forgotten). In early phases of policy-making, this option is often to be preferred. However, it is of the utmost importance that any form of self-regulation is assessed with special care and is verifiable, transparent to public authorities and subject to periodic reviews and evaluations. Very often, in the AI-enabled environment, this option can be coupled with automatic data exchanges, similar to the ones implemented under RegTech and SupTech.
    - *Bottom-up (civil society-driven) options*. Rather than relying on self-regulation, one slightly different set of alternatives could be, to rely on third party regulatory schemes, privately developed by civil society organizations or NGOs, and use them as private standards. For example, the ISEAL alliance acts as a private meta-regulatory scheme, which defines rules and criteria that private regulators should follow in order to be credible interlocutors of public policy-makers (Cafaggi and Renda 2014); in the internet environment, an NGO (or public authority when feasible) could define or adopt criteria for algorithmic accountability, transparency and non-discrimination, and translate them into an auditing and certification scheme, which ultimately

[75] On the use of cost-benefit analysis in EU Better Regulation and possible shortcomings of this method, particularly when it comes to non-market impacts (including impacts on fundamental rights) and distributional impacts, see Renda (2018b).

leads to a trust-enhancing tool for end users. A specific logo/label could be used for those platforms whose algorithms have been checked and are constantly monitored by public authorities or NGOs, so that compliance is ensured over time.

- *Co-regulation*. Co-regulation couples the delegation of specific phases of the policy cycle to the private sector, with a legal backstop. In the case of AI, co-regulatory schemes could be used whenever monitored self-regulation is (or has already proven to be) ineffective in securing an adequate level of protection of users' rights, or any other lack of effectiveness in achieving policy objectives.
- *Principles-based and outcome-based regulation*. Very often, the performance of specific phases of the policy cycle by private entities can be coupled with the identification of the main overarching principles of the regulatory intervention, and/or the ultimate goals the policy intends to pursue. These forms of regulation often achieve greater flexibility compared to traditional, "command and control" regulation, since they can be coupled with the adoption of more agile secondary legislation, such as implementing or delegated acts, or measures taken by a regulatory agency; or with private regulatory solutions. In either case, the resulting policy framework can be adapted more easily to the fast-changing technological evolution. The most evident case of flexible regulation in this respect is the EU's "new approach to standardisation", in which principles and objectives are set in primary legislation, but all implementing measures are then achieved through standards and conformity assessment. For example, recent laws on net neutrality and on copyright infringement mandate that internet service provider (ISPs) act as controllers of the behaviour of their subscribers. Online advertising, child protection and privacy are still subject to a combination of general legislation and private standards in many parts of the world. The internet itself is subject mostly to private regulation. Table 1 below shows Chris Marsden's "Beaufort scale" of the many hybrids that exist in the ICT ecosystem between pure self-regulation and full-fledged co-regulation.[76]

[76] A Beaufort scale is originally a scale for measuring wind speeds: Chris Marsden used this denomination to show a taxonomy of forms of self- and co-regulation ordered from the least managed, to the most government-imposed form.

Table 1. Marsden's "Beaufort scale" of self- and co-regulation

Scale	Regulatory scheme	Self-Co	Government involvement
0	'Pure' unenforced self-regulation	Second Life	Informal interchange only – evolving partial industry forum building on players' own terms
1	Acknowledged self-regulation	ATVOD	Discussion but not formal recognition/approval
2	*Post-facto* standardized self-regulation	W3C#	Later approval of standards
3	Standardised self-regulation	IETF	Formal approval of standards
4	Discussed self-regulation	IMCB	Prior principled informal discussion, but no sanction/approval/process audit
5	Recognised self-regulation	ISPA	Recognition of body – informal policy role
6	Co-founded self-regulation	FOSI#	Prior negotiation of body – no outcome role
7	Sanctioned self-regulation	PEGI# Euro mobile	Recognition of body – formal policy role (contact committee/process)
8	Approved self-regulation	Hotline	Prior principled less formal discussion with government – with recognition/ approval
9	Approved compulsory co-regulation	KJM# ICANN	Prior principled discussion with government – with sanction/ approval/process audit
10	Scrutinised co-regulation	NICAM#	As 9, with annual budget/ process approval
11	Independent body (with stakeholder forum)	ICSTIS#	Government imposed and co-regulated with taxation/compulsory levy

Note: # denotes 'soft power' of government/European Commission funding.
Source: Marsden (2011).

As mentioned also in Cafaggi and Renda (2012), regulators could find it useful to break down the phases of the regulatory intervention (e.g. into agenda-setting, standard-setting, implementing acts, monitoring and evaluation, enforcement) and decide which phases would be more effectively dealt with by private actors, as opposed to others that should remain within the remit of public authorities. In the case of algorithmic awareness, co-regulation could mostly take the form of either: i) principles-based regulation, in which public legislation sets the principles that have to be followed by online platforms in using algorithms, and then platforms complement the rules by developing compliant business models and algorithms and forms of reporting to public authorities that enable monitoring and evaluation over time; or ii) outcome-based regulation, in which public authorities directly specify outcomes and performance indicators that should be complied with by the regulated platforms (e.g. using KPIs (key performance indicators) and coupling them with a successful monitoring strategy).

- *Behavioural (choice architecture, hyper-nudge) options*. These options factor behavioural biases into the analysis and seek to preserve the choice of the end user by, at the same time, gently pushing them to adopt solutions that are considered more sustainable, or more in line with end-user rights. There is still a lively debate on the merits of nudging as a truly libertarian approach (Cass Sunstein speaks of "libertarian paternalism" claiming that it is not an oxymoron); in the context of cyberspace, and even more in the case of algorithms used by online platforms, mandating the use of certain default options (e.g. opt-in schemes for personal data collection), or the use of specific colours or other graphical tools to increase the likelihood that users will choose more conservative or sustainable options. Issues such as the endowment effect, over-optimism bias, short-termism, anchoring and framing are essential to assessing ex ante the likelihood that a policy approach will be effective in tackling the issue.
- *Design-based (code-based, architectural) options*. Some policy approaches entail the partial avoidance of awareness-raising, and instead focus on mandating that the overall architecture of the platform is not conducive to infringing conduct or undue discrimination. This approach dates back to the early literature on cyberspace, when Lawrence Lessig (1999) spoke of "the perfect technology of justice"; more recently, Yeung and Dixon-Woods (2010) stated that "design-based regulation … works ex ante: it uses technical constraints to stop, or significantly inhibit, action at the moment it is attempted". Hence, this approach avoids the risks, uncertainties and inefficiencies generated by ex post remedies, as well as the risks of behavioural regulation (such as nudges, see below), by avoiding the emergence of specific discriminatory outcomes.
- *RegTech* and *SupTech* solutions, currently confined to the financial services sector, but potentially applicable also to many other sectors. Technology can be used in any regulated sector to conduct risk assessments, monitor, report and comply with regulatory obligations at reduced cost. In addition to digitalisation of reporting, technology can carry out automated compliance processes and checks, it can reduce operational risk, increase cybersecurity, and provide real-time analysis and compliance ("Compliance as a

Service"), which in turn can be used to combat fraud and other crimes, as well as to provide warnings on a range of other issues. Technology can include digital labour, robotic process automation, machine leaning, cognitive learning, big and smart data analysis, biometric technology, and natural language processing. RegTech does not necessarily clash with the need for human input to update regulations and also to buy-in to the culture of regulatory compliance.

- *Adaptive, experimental regulatory options (experimental policy-making).* The speed at which technology progresses and the disruptive nature of new business models in many sectors led regulators around the world to seek more experimental approaches to regulation, which in turn enable better monitoring and evaluation over time. Possible ways to make policy more flexible and adaptive include: i) the use of *regulatory sandboxes* and other experimental approaches to allow for the ongoing monitoring of the market and social impacts of innovative techniques; ii) the *incorporation of technology roadmaps* and the opinion of multi-stakeholder platforms as input into the policy-making process, to ensure that innovative, welfare-enhancing technologies are adequately represented in policy processes and outcomes; and iii) the *ongoing monitoring* of policy impacts, including through open government technique.[77] This leads to an enhanced role of certain better regulation tools, such as the use of sunset clauses and forms of experimental policymaking, to trigger learning on the side of government (Listokin 2008; Ranchordás 2013). The "adaptive regulator" would then be guided by a number of principles, including: an *incremental* approach, meaning that small steps should be taken and social change should be based on experience; an *experimental* approach, justified by the "combination of uncertainty and constraints on predictability [which] create … the necessity for policymakers to experiment;" and a *flexible* approach, required by the existence of deep uncertainty. This is where this literature stops, and where new insights would be urgently needed. Yeung (2017b)

[77] In academia, very few commentators have directly addressed the issue of flexible, adaptive policy-making in the layered ICT ecosystem. In their attempt to propose an adaptive framework for the internet, Clark and Claffy (2015) argue that the following requirements are essential: agreeing on policy goals; measuring progress towards those goals; designing regulatory options intended to move towards those goals; being able to determine that policy changes indeed caused observed outcomes; and dealing with the potential destabilization of ecosystem, due to rapid policy and technology adjustments. Likewise, policy-makers should possess the flexibility to revise and adapt the structure of policies and programmes to changing circumstances.

provides a classification of so-called "fixed" and "adaptive" regulatory approaches to algorithms, as shown in Table 2 below.

Table 2. A taxonomy of algorithmic regulatory systems

	Standard setting	**Monitoring**	**Enforcement/ Sanction**	**Description**
1.	Fixed	Real time reactive violation detection	Automated	Simple real-time sanction administration systems
2.	Fixed	Real time reactive violation detection	Recommender system	Simple real-time warning systems
3.	Fixed	Pre-emptive violation prediction	Automated	Simple pre-emptive sanction administration systems
4.	Fixed	Pre-emptive violation prediction	Recommender system	Simple predictive recommender systems
5.	Adaptive	Real time reactive violation detection	Automated	Complex sanction administration systems
6.	Adaptive	Real time reactive violation detection	Recommender system	Complex real-time prioritization systems
7.	Adaptive	Pre-emptive violation prediction	Automated	Complex predictive sanctioning systems
8.	Adaptive	Pre-emptive violation prediction	Recommender system	Complex predictive recommender systems

Source: Yeung (2017b).

- *Command and control options.* These correspond to standard prescriptive policy options, possibly coupled with review clauses that could increase the flexibility of the rules over time. Banning specific forms of behaviour, for example profiling of end users, requires an adequate strategy for monitoring compliance, which in the past has been missing. Accordingly, in comparing these prescriptive options with more flexible approaches, we will guide policy-makers not only on the basis of the design and text of the rules, but also, most notably, in the appraisal of the risk of low compliance. A good example is again p2p (peer-to-peer) file sharing, where technology has constantly outpaced regulation in developing solutions and user-friendly opportunities to achieve non-compliant outcomes (Renda 2011).

- **Risk analysis should become far more embedded in the regulatory practice of the European Commission.** The analysis of the factors that may jeopardize the effectiveness of policy solutions, and the identification of possible mitigating strategies, appears inevitable in a future that is likely to feature more "code as law" solutions. As Lessig himself observed in 1999, and as was confirmed by past experience (e.g. in copyright and DRMs), regulation by means of code is often unable to represent the level of flexibility and the nuances required for adequate interpretation of specific provisions (e.g. copyright exceptions).

- **In the areas of monitoring and evaluation**, any alternative policy framework should be *adaptive* and feature a strategy for data collection in order to enable monitoring and possible changes over time. In the case of co-regulatory or monitored self-regulatory schemes, the private sector should cooperate with the EU institutions in order to enable seamless monitoring of existing policy solutions. In the case of RegTech (Regulatory Technology) solutions, such cooperation would be automatic, due to the partial replacement of legal rules with computer code (Micheler and Whaley 2018).

5.2 Reforming liability rules?

One key aspect of the future policy framework for artificial intelligence is the choice of the liability regime for damages caused by AI systems: some member states and the European Commission have launched reflections in this domain.[78] This is particularly relevant in case of damages suffered by end users (in B2B cases, liability is typically a contractual issue, rather than a tort one). There are three main aspects of this issue, which will have to be discussed and decided upon in the coming months. The first is related to the *scope* of the liability; the second to the *type of remedy*, and hence the type of liability rule to adopt; the third revolves around problems of *attribution or apportionment* of liability. Concerning the *scope*, it is essential to discuss whether developers, vendors or distributors of AI systems should be liable for damages caused by the selection of the data; for the way in which they clean or otherwise curate the data; for failure to adopt adequate safeguards when training machines; for damages caused by the design of the algorithm; and/or for damages caused irrespectively of the respect of all standards of conduct in these phases. It is useful, in this respect, to distinguish between input, throughput, output or impact accountability.

[78] E.g. the UK government has conducted detailed analysis of the applicable regulatory regime and produced a series of reports, including *The pathway to driverless cars: a detailed review of regulations for automated vehicle technologies*, Department for Transport, 2015. The European Commission published a Staff Working Document on "Liability for emerging digital technologies", on 25 April 2018. For a review of the issues surrounding new regulating technologies, see the report of the RoboLaw Project, *Guidelines on Regulating Robotics*, 22 September 2014.

Secondly, the *type* of liability rule to be adopted can take the form of a fault-based rule, in which the liability of the alleged tortfeasor is gauged with respect to a standard of diligence, which can be well-defined (for example, associated with compliance with a specific standard); or more subject to judicial interpretation (e.g. the diligence of the layperson, or good *pater familias*, or the diligence and competence of an expert in the field).

Thirdly, and in a related vein, it is important to clarify the rules that apply in case of difficulty to attribute the responsibility to a given AI system. This can happen due to any of the following scenarios: i) a system good causes a given damage, but the individual contribution of AI to the damage is impossible to prove; ii) an AI system did not incur any malfunctioning, but its interaction with human behaviour led to damage; iii) an interaction between two or more AI-enabled algorithms has caused damages to third parties (e.g. so-called "flash crashes"); iv) the combination of two or more AI systems, from different vendors, within a single product leads to damages, with no easy apportionment of liability between the system vendors; or v) it is difficult to prove who, between the AI vendor, the distributor, or the OEM (original equipment manufacturer), has caused the damage.

For example, in the case of the fatal accident that occurred in March 2018 in Tempe, Arizona, when a Uber-operated Volvo car failed to detect a woman who was crossing the street, public authorities took several days and had to closely cooperate with Uber to trace back responsibility for what had happened. Was it the Lidar sensor, and then its producer should be liable? Was it a mechanical failure, and then Volvo should be liable? Was it the camera? Was it Uber, who runs the operating system of those cars? The NTSB Preliminary Report indicated that Uber had deactivated at least two safety-critical features, including the emergency braking. But there were also concerns that the "human in the loop" was watching a TV show on her phone rather than being ready to step in; but her declarations also raised issues on possible lack of training, which would cause liability to shift back to Uber;[79] moreover, the victim was reportedly under the influence of drugs and alcohol, and it is unclear whether this could have affected the predictability of her behaviour, or created issues of

[79] The Preliminary Joint Consultation Paper on Automated Vehicles, by the Law Commission of England and Wales, and the Scottish Law Commission, examines the issue of negligence by a "user-in-charge". See https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2018/11/6.5066_LC_AV-Consultation-Paper-5-November_061118_WEB-1.pdf

contributory negligence (see below). This suggests how burdensome fact-finding can be in complex situations in which several factors cause an accident.[80]

In this context, the law and economics literature can be of great help. Scholars such as Guido Calabresi have shed light on the most efficient choice of the liability regime. Key criteria in designing an efficient liability regime are the identification of the "cheapest cost avoider", i.e. the entity that can avoid the emergence of system failures at the lowest cost; and /or the "superior insurer", i.e. the entity that can buy insurance most effectively and cheaply, thereby offering relief to the damaged party also due to voluntary or mandatory insurance (so-called "deep pocket theory"; see Posner and Landes 1980). The latter concept also leads to another, very important issue to be included in the overall discussion of liability for use of AI: the need to ensure that damaged victims are always adequately compensated for damages. This is essential, regardless of the fault of the company that designed the AI system, provided that the end user has not used it incorrectly.

As a matter of fact, the acceptability of new disruptive technologies such as AI is dependent on the fact that end users damaged by the system find appropriate redress. Furthermore, the need to identify a responsible entity for damages caused by AI is linked to the introduction of a related principle in the Ethics Guidelines, which we discussed in section 4 above: the "human in the loop" or the "human in control" requirement should rather be replaced by a "human responsible" requirement, to avoid imposing excessively burdensome obligations on AI developers, vendors and distributors in circumstances in which it is virtually impossible or useless to have a human in the immediate control of the system; and at the same time guarantee that end users will be compensated for the damage caused, and will therefore be more likely to accept and take up the new systems.

This approach to responsibility inevitably leads to the identification of a strict (non-fault-based) liability regime. **However, this would not, in and of itself, be sufficient for the design of a suitable liability regime**. In particular, three aspects would need to be defined in detail: whether the regime would be absolute or relative; whether there would be one entity in the whole value chain that is primarily responsible vis-à-vis the end user; and whether there would be joint and several liability in case of joint participation in causing an accident. Concerning the extent of the liability regime, strict liability regimes can be absolute, or relative. In the former case, there is no possibility to escape liability:

[80] Another, less recent example of causative uncertainty is the *Bookout/Schwarz* litigation around unintended acceleration in Toyota vehicles. In that case, the NHTSA struggled to establish the proximate cause of the accident and had to instruct NASA to investigate. After 10 months, NASA failed to form a definitive view about causation.

the mere existence of a damage, and a causal relationship between actions taken by an AI system and the damage, triggers liability and then damage compensation (or specific performance). A good example of cases of absolute strict liability is the employer's liability for acts of their employees while in the execution of their duties. Relative strict liability allows for a broader set of exceptions, often related to the foreseeability of damages at the time the product was placed on the market; or the fact that the alleged tortfeasor took all reasonable measures to avoid the damage, and that the latter resulted from events that were beyond the alleged tortfeasor's sphere of control. A typical case of relative strict liability is the current regime for product liability, but also liability for damages caused by animals, things under custody or in the exercise of so-called "dangerous activities" (including the act of driving).

With respect to liability vis-à-vis the end user, many legal systems have tried to facilitate victims of torts or breach of contract, especially when they are consumers, by creating a "one-stop-shop", i.e. a single entity that is responsible for consumer redress. In the case of product liability, it is the producer itself that is strictly liable towards consumers. In the case of contractual guarantees in the sale of goods, the vendor is liable. In all those cases, the "one-stop-shop" allows the first responsible entity to sue other entities in the value chain to obtain redress. For example, if an accident is caused by the lidar sensor producer of a self-driving car, the car operator (e.g. Waymo, or Uber) would be liable towards damaged third parties; but they would then have the chance to seek redress from the lidar sensor producer to recover part or all the money paid to compensate for damages.

Moreover, another issue to be discussed is the possibility that the extent of the liability would be mitigated by the fact that end users have misused the product, or in any way taken insufficient care while using an AI-enabled system. As suggested i.a. by Cooter and Ulen (2004), coupling strict liability with contributory negligence can provide optimal incentives for both parties involved in a dangerous activity. However, end users have to be provided with sufficient information and advice on how to handle a given AI product, or otherwise the likelihood of unintentional misuse of AI systems would increase, often leaving end users without coverage for damages.

Finally, in the case of a "flash crash", or all other cases in which damage is caused by the interaction between algorithms and the external environment, including other algorithms, it may be difficult to apportion liability among two or more entities. In these cases, legal systems often have no answer to the question "who is responsible, and for how much?" For similar problems, some legal systems have foreseen cases of joint and several liability, in which each of the parties is responsible for the entire damage caused but can then sue the other parties to obtain partial compensation.

5.2.1 *Objects, animals, slaves or robots?*

The design of a liability regime for AI inevitably boils down to a fundamental question: Can AI be considered as an object under the control of a human being, or does AI feature some elements of autonomy, which would warrant a different set of rules? In the previous section, our discussion referred to actions of human beings and the potential defectiveness of objects. In this section, we broaden the discussion to a number of possible alternatives.

First, if AI is considered as an extension of the human being, or a part thereof (as could occur in the case of augmented intelligence), then the liability rules applicable to humans would also apply to the AI system. Accordingly, a fault-based regime will most often apply: in many civil law countries, such rule will go back to the Roman *lex aquilia*, which requires a subjective element (negligence, or the intention to cause harm), an unjust damage being caused to another party, and a causal connection between the two.

Second, if AI is considered as equivalent to an object, then the so-called *res ipsa loquitur* (also a common law doctrine) could apply: under this rule, negligence can be presumed if one's property causes harm to a third party. But where no negligence is found on the part of the custodian, owner, or user, liability can be transferred to the manufacturer of the AI-enabled system. This, in turn, will lead to problems of apportionment of liability, as mentioned above, and recently reiterated by Giuffrida et al. (2018).[81] The alternative approach to *res ipsa loquitur*, as discussed above, would be outright no-fault (strict) liability, which is construed by some scholars also as a fault-based system, configuring a duty to exercise care in monitoring objects under custody (*culpa in vigilando*).

Third, it is reasonable to expect that AI will be used mostly "as a service", especially by SMEs. In that case, it would not be a product but a service that causes damages. In those circumstances, an open question is whether the resulting responsibility for damage caused by an AI system should be of a contractual nature (i.e. provision of a service that does not conform to sufficient security requirements), which does not exonerate the purchasing party from

[81] Giuffrida et al. (2018) also quote the Florida Statute Fla. Stat. § 316.86 (2016) exempting automobile manufacturers from liability when third-party AI is installed: "The original manufacturer of a vehicle converted by a third party into an autonomous vehicle is not liable in, and shall have a defense to and be dismissed from, any legal action brought against the original manufacturer by any person injured due to an alleged vehicle defect caused by the conversion of the vehicle, or by equipment installed by the converter, unless the alleged defect was present in the vehicle as originally manufactured.").

liability towards damaged parties; or of a non-contractual nature (tort liability), which would then have to be extended to services.

Fourth, an AI system could be considered as similar to an animal, especially when it displays a certain degree of autonomy. This option is possibly backed by authoritative statements in the AI field, which compare the intelligence of most advanced AI systems to that of a small animal, like a frog or a cat. This option would also imply that AI systems have no legal personhood, and that strict liability applies only in case of damages caused by dangerous animals, such as wild animals, if they were not duly kept under custody. A similar rule exists both in civil law countries and in the US common law system.[82]

Fifth, AI could be considered as a "slave". This interpretation is backed by the fact that the word "robot", in its original Czech word, means "forced labour" or "slave". Soluim (1992) and Hubbard (2011) discuss this option. In Roman law, masters were liable for damages caused by their slaves. And in the United States, a master was liable for every [slave's] trespass, whether the act was done when in the master's service, or not, and whether it was done with or without the master's knowledge.

Sixth, AI could be considered as an employee, and be given legal personhood as well as the duty to exercise due care. Strict liability would still be attributed to their owners, but the AI system would be given legal personhood and could, in principle, be asked to compensate for the damage. This perspective appears to be deeply related to the belief that AI systems may display, in the future, a significant degree of autonomy with respect to their "owners" (developers, trainers, programmers, vendors). Recent breakthroughs in AI, mostly due to the use of Deep Learning and Deep Reinforcement Learning techniques, are first steps towards distancing the acts of the AI system from the will of the programmer. At this stage, however, postulating (like the European Parliament did in 2016) smart autonomous robots with rights and duties seems to be at least premature and would also lead to a situation in which no certainty is given to damaged parties as to who should, and will, compensate the damage. The same could be said about an even more extreme scenario depicted by the European Parliament: a situation in which AI systems (and in particular, robots) are not considered as employees, but as outright legal persons, with no link to an "owner" or developer.

All in all, the choice between these options should be dictated by a discussion of the reality of AI, rather than its associated myths; by the need to

[82] *Behrens v. Bertram Mills Circus*, Ltd. [1957] 2 QB 1, 11 (Eng.). The acts of wild animals give rise to strict liability. Others, especially domestic animals, impose tort liability only if harm is foreseeable.

ensure that victims of actions carried out or inspired by AI systems obtain adequate compensation; by the need to avoid stifling innovation by expanding liability to unchartered territories, beyond what is reasonably foreseeable at the time of AI development and commercialisation; and by the need to ensure that humans remain at the centre of both legal rules and AI development. **In all these respects, the current EU legal framework appears largely adequate, but may need some clarification and interpretive guidance in order to avoid generating confusion and a lack of certainty among industry players.**

5.2.2 Current EU law

The current EU rules on liability for AI systems is mostly related to two pieces of legislation: the Product Liability Directive and the Machinery Directive.[83] The former Directive foresees that if a product causes damage to a person or their private property, the producer is liable to pay compensation. A recent evaluation showed that the Directive continues to strike a good balance between consumer protection and encouraging innovation in the EU.[84] The Commission, however, has promised to publish interpretative guidance by mid-2019 to facilitate a common understanding among EU countries and to further clarify to what extent it applies to emerging technologies. The Commission announced that, if necessary, it will update certain aspects of the Directive, such as the concepts of 'defect', 'damage', 'product' and 'producer'. In this respect, key aspects will have to be clarified, including: i) how to interpret the foreseeability of damage; ii) how to construe the so-called "state-of-the-art" exception to liability; iii) whether to include "as a service" use of AI within the scope of the Directive; as well as to (iv) how to ensure that the definition of misuse of an AI product does not place too much risk on the side of the end users.

On the other hand, the Machinery Directive sets general health and safety requirements for products, such as robots or 3D printers. This is a good example of outcome-based legislation, in line with the "new approach": the choice of technical solutions that should be deployed to meet the high safety level is then left to manufacturers. Here again, whether safety levels should be related to the performance of the product in real life (e.g. including interaction with humans,

[83] Directive on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products, 85/374/EEC, OJ L210, 07/08/1985, pp. 29-33. And Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery, and amending Directive 95/16/EC (recast) OJ L 157, 9.6.2006.

[84] For both evaluation reports, see https://ec.europa.eu/luxembourg/news/commission-publishes-evaluation-reports-eu-rules-machinery-safety-and-product-liability_fr.

or other algorithms) or in a lab, should be subject to clarification. And possibly, the Machinery Directive is one of those pieces of legislation for which regulatory sandboxes, including virtual sandboxes, could be foreseen in order to ensure that the testing of new machinery takes place in a real (or accurately simulated) environment, for a sufficient amount of time to allow for testing.[85] The European Commission has already committed to launch a study to further look into certain aspects of emerging technologies, such as issues arising from human-machine collaboration, which are not explicitly addressed by the Directive.

Last, the future EU liability regime will also have to be designed in combination with a suitable insurance framework. Assuming that it is always possible to trace back a liable entity/person in an automated process, the future Ethics Guidelines could require that any company developing, embodying or selling AI in their systems checks their financial ability to respond to any potential liabilities that could arise from its use. In case that is not possible, users should be required to abandon that use or to cover those risks with an insurance policy or an equivalent requirement. Should the insurance system end up being too burdensome, especially for SMEs, a mandatory, subsidized insurance system could be foreseen, in order to combine the benefits of innovation with the certainty of compensation for end users.

5.2.3 *Avoiding overlapping regulatory requirements in heavily regulated sectors: Towards a "tech REFIT" strategy?*

In the previous sections, we have approached the issue of AI regulation from a cross-cutting perspective, looking at various approaches to regulation, as well as horizontal pieces of regulation such as the GDPR, the Product Liability Directive and the Machinery Directive. At the same time, there are many pieces of sectoral regulation that already impose specific behaviour on the side of producers and service providers, which should not be cumulated with new regulatory obligations, in order to avoid redundancies and overlaps, with consequent losses of legal certainty and productivity.

For example, insurance companies and banks are already subject to several information disclosure requirements, which may easily **overlap with future regulatory obligations of transparency, accountability and non-discrimination in future AI policy**. The Insurance Distribution Directive (Directive (EU) 2016/97), implemented as of 1 October 2018, is a good example

[85] As defined by the UK Financial Conduct Authority, a virtual sandbox is "an environment that enables firms to test their products and services in a virtual space without entering the real market (for example, by testing with publicly available data sets, or with data provided by other firms through the virtual sandbox)". See https://www.fca.org.uk/publication/research/regulatory-sandbox.pdf.

as it requires i.a. that, prior to the conclusion of a contract, distributors specify the needs of customers and provide certain objective information on this basis to allow the customer to make an informed decision. Similarly, MiFID II created the need to record large amounts of data that are well-defined and structured for regulatory review and sharing across counterparties and trading venues, paving the way for a large RegTech exercise. Firms will be reporting various pre- and post-trading data, but also venue of execution, venue of publications, transaction ID code and much more. Other provisions of this sort were introduced in other regulated sectors, from healthcare to e-communications, energy and transport (see Box 3 below on automated driving).

While a full analysis of all these overlaps would fall outside the scope of this report, it is important to reflect on the need for a thorough review of the existing legislation before introducing new obligations across the board. **One possibility would be to work in the direction of a "tech REFIT", i.e. an expansion of the ex-post evaluation methodology, aimed at introducing specific questions as regards the compatibility between existing legislation and the possible, future policies on AI** transparency and accountability, as well as non-discrimination emerging from the Ethics Guidelines. Such a tech REFIT methodology would be a relatively straightforward addition to the current Better Regulation guidelines, as well as to the work of the REFIT platform set up in 2015 by the European Commission.[86]

Box 3. The new UK law on automated and electric vehicles

The UK has recently addressed some of the emerging issues in liability for AI systems in the Automated and Electric Vehicles Act, adopted in July 2018. The 2018 Act features a 'light touch' approach and is expected to be updated as the technology and infrastructure progress. This was the first piece of legislation worldwide to set down a formalized legal model for the insurance and liability of automated vehicles. Not surprisingly, technological uncertainty weighs heavily on the quality of the drafting, and as such should lead EU institutions and other governments to use caution in referring to it as a model (Channon et al. 2019). It should also be borne in mind that the 2018 Act still requires further secondary legislation to bring it into force.[87] However, it is interesting to report at least some of the basic feature of this law.

[86] https://ec.europa.eu/info/law/law-making-process/evaluating-and-improving-existing-laws/refit-making-eu-law-simpler-and-less-costly/refit-platform_en

[87] Meanwhile, the UK government initiated a three-year review of the law on regulating Automated Vehicles in 2018. To this end, the Law Commissions of England and Wales, and

The Act provides the possibility for insurers to limit their liability under section 2(1) to their policyholders and, in certain narrow circumstances, to third parties whenever accidents were caused by the failure of an insured person, or with the insured person's knowledge, to install safety critical software updates. Likewise, accidents caused by software alterations prohibited under the terms of the vehicle insurance policy may also be excluded from the coverage. Nonetheless, insurers will still be strictly liable to compensate the majority of third parties injured in accidents to which the 2018 Act applies.[88] This has significant implications, as it effectively gives insurers a role in policing autonomous vehicles' software. Accordingly, over time, insurers will end up putting pressure on AI providers to move towards some level of certification, as insurers cannot be expected to maintain a rolling review of all autonomous vehicle software (unless some form of automated data sharing and monitoring is foreseen). The Act also effectively creates a statutory right of subrogation for insurers/owners to recover their losses in full or part from any other party responsible for the accident. Thereafter, it is anticipated that insurers will pursue the OEMs of the autonomous vehicle for recovery of their outlays under the existing Product Liability Regime. This may lead insurers to advance claims under contract and in negligence under tort law.

Concerning complementarity between man and machine, a recent Government Code of Practice for testing autonomous vehicles clarified that a test driver and/or operator will need to be able to assume control of the vehicle at any time.[89] The Code also envisages scenarios in which remote operation of an autonomous vehicle will be necessary. As such, it appears that an 'in the loop' connection will still be required in the short to medium term. Such role may even be played by insurers, who could provide real-time input/risk assessments into autonomous vehicles driving.[90] Nevertheless, it remains to be seen whether insurers will want to be exposed to this level of risk.

Scotland, have produced their first Preliminary Consultation Paper. Further papers will follow in due course. The Paper is detailed, extensive and the result of sustained engagement with a broad range of stakeholders. It is likely that the Consultation responses will lead to significant improvements to the 2018 Act.

[88] This is because the Road Traffic Act 1988 remains in force. The 1988 Act has been amended so that insurance policies required under s 145 must also provide for an insurer's obligations under s 2(1) of the Automated and Electric Vehicles Act 2018.

[89] See UK Department of Transport, Pathways to Driverless Cars: A Code of Practice for Testing (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/446316/pathway-driverless-cars.pdf).

[90] See https://www.technologyreview.com/s/611003/one-way-to-get-self-driving-cars-on-the-road-faster-let-insurers-control-them/

5.3 Boosting data-driven innovation through ad-hoc policies?

Perhaps the area in which the EU institutions should be more proactive in the years to come is that of data-driven innovation. As a matter of fact, as was remarked in several occasions during the meetings of the CEPS Task Force, **most often the opportunities and the challenges that are attributed to AI development are more easily and appropriately referred to as data** and the use of large datasets in the emerging technology stack. The lack of data has emerged as a key issue for many entrepreneurs, large and small companies, willing to use machine-learning techniques. And while it is true that not all AI is as data-hungry as machine learning, it is important to stress that most of the emerging, disruptive uses of AI in several policy fields rely on some variant of machine-learning techniques. This is also a problem when it comes to international competitiveness: many experts are of the opinion that countries like China and the United States will be able to rely on greater data availability than Europe, also (but not exclusively, and probably not primarily) due to the more restrictive privacy laws in force in European countries.

To be sure, **data are more accessible today than in the past: but not all players in the market can have access to the same amount of data, and this may stifle the competitive dynamics in specific markets or create collective action problems in others**. So far, however, the academic literature in either economics or in industrial organisation has found data to be a stand-alone entry barrier for companies wishing to start competing with incumbent players. And indeed, imposing data-sharing across the board for all players in the market would probably backfire, since incentives to invest in the production and analysis of data would be inevitably weakened by the perspective of having to share, most often at regulated (FRAND) prices, the whole dataset. Very often, proposals that aim at imposing data-sharing obligations are based on the assumption that data are a non-rivalrous good, i.e. its value does not change according to how widely it is shared. From a dynamic efficiency perspective, however, data are rivalrous, since the overall investment in their collection, production and elaboration depends on how much the involved players expect to profit from the activity. Following Bebchuk (2001), the choice between a property and a liability rule appears very different if seen from an ex-ante perspective.

How can the right balance be struck between the need to promote data availability, and the need to maintain proper incentives with respect to large datasets? The CEPS Task Force has converged over the following recommendations:

- **It is of the utmost importance that governments adopt *open data* policies, by making available to the public large datasets, possibly in formats that are interoperable with existing machine-learning software**. So far, data held by government and data from publicly funded research are still largely unavailable to researchers, entrepreneurs and companies willing to engage in data-driven innovation. In this respect, following the presentation of the European Commission's Data Package on 25 April 2018, a revision of the Public Sector Information Directive has been tabled.[91] The PSI Directive first came into effect in 2003, and was amended in 2013 to clarify that i) PSI should be presumed to be "reusable by default," ii) museums, archives and libraries were subject to the Directive's provision, iii) acquisition fees were limited to marginal costs of reproduction and iv) documents were to be made available for reuse using open standards and machine-readable formats. The ongoing review could lead to further changes to improve reuse of public sector information, with reference to the available data, their accessibility through API (application programming interface) and the applicable economic conditions.
- **The free flow of data in the Single Market should be further promoted**, in line with the ambition of the Regulation adopted in November 2018. At the same time, the possible exceptions to the free flow, for example, based on national security stances, should be narrowly interpreted to avoid disproportionate disruptions of data flows. The Regulation prohibits data localisation restrictions, thereby permitting organisations to store data anywhere in the EU. It allows competent authorities to access data – for scrutiny and supervisory control –regardless of where it is stored and/or processed in the EU. The Regulation encourages the creation of codes of conduct for service providers who process data (for example, cloud service providers) in order to facilitate switching between providers in a structured and transparent manner.
- In order to reconcile data availability with the need for data protection, the **European Commission should fund research, innovation and standardisation in the domain of privacy "by design", as well as privacy-enhancing technologies**. In the case of large datasets, key technologies include cryptographic solutions that allow for the use of large datasets without infringing privacy laws, such as private practical computation, zero-

[91] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and The Committee of the Regions, "Towards a common European data space" COM/2018/232 final, 25.04.2018.

knowledge proofs[92] or homomorphic encryption.[93] These techniques are easily coupled with sectoral, blockchain-based initiatives aimed at sharing data among competitors or complementors along the same value chain.

- **Allowing text and data mining for both research and commercial purposes would be very important for data-driven innovation in Europe**. This topic is of course very controversial in Europe, as demonstrated by the recent, heated debate on copyright reform. Data mining, however, is essential for the future competitiveness of the EU: if subject to ethical guidelines (so-called "Ethical Data Mining"), it could become an engine of data-driven innovation in Europe.[94]
- **Experimentation is key to keeping Europe relevant in the AI field**. However, the **GDPR's data minimization principle and the need to have a clear purpose for getting user consent can limit the ability to experiment** with innovative approaches, even when users have given explicit consent to access their data.
- At this stage **it seems wise to avoid more strictly regulating access to data,** in particular outside the rather confined remit of competition law and refusals to deal by dominant companies. The recent Communication "Towards a Common European Data Space" (EC 2018) argues that in general stakeholders do not favour a new data ownership type of right and indicate that the crucial question in business-to-business sharing is not so much about ownership, but about how access is organised.
- There is a **need for more clarity on the legal framework for machine-generated data**, as had been mentioned in the Commission's Communication on Building a European Data Economy. Current proposals in the package on data of April 2018 focus on the review of the PSI Directive for data held by the public sector (including public undertakings) and on soft law for access to and preservation of scientific information. As for access to and re-use of private sector machine-generated data in business-to-

[92] Zero-knowledge proofs (ZKP) are advanced cryptographic techniques that allow someone to produce proof of a statement without disclosing the data underlying that statement.

[93] Advanced cryptographic methods that allow someone to request distributed computations to be performed by private servers. While the underlying data of these computations are never revealed or shared on the blockchain, it is theoretically possible to obtain a cryptographic proof that the aggregated result of these computations is correct. these techniques would be implemented outside of the blockchain network ('off-chain') but it could potentially be useful to use the blockchain to store these proofs of computation for every stakeholder to see.

[94] See Report of the Expert Group chaired by Ian Hargreaves (http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf).

business relations, the Communication "Towards a Common European Data Space" defines a series of key principles that should be respected in contractual agreements in order to ensure fair and competitive markets. These principles include:

i. transparency, i.e. transparent identification of the entities that will have access to the data, the type of such data and at which level of detail and the purposes for using such data;

ii. shared value creation, i.e. acknowledgement that where data is generated as a by-product of using a product or service, several parties have contributed to creating the data;

iii. respect for each other's commercial interests and secrets of data holders and users;

iv. undistorted competition when exchanging commercially sensitive data;

v. minimal data lock-in, i.e. enabling data portability as much as possible.

Separate key principles are proposed by the Commission for data sharing in business-to-government relations, including among others proportionality in the use of private sector data and purpose limitation.

More generally, at this stage it seems wise to avoid regulating access to data more strictly, in particular outside the rather confined remit of competition law and refusals to deal by dominant companies (Drexl 2017a and 2017b).

- **In specific sectors, the issue of data sharing is becoming a reality**, and the European Commission seems increasingly determined to open some markets to competition by forcing incumbent firms to share their data with new entrants. This approach is also echoed by scholars who would want to see data accumulation as a case of essential facilities, as such potentially conducive to compulsory licensing when imposed by a dominant firm, and under certain conditions (Graef 2016; Renda 2010). However, in some cases these initiatives appear to create even bigger problems than the ones they are meant to solve. For example, in the case of the revision of the Payment Services Directive (PSD2), imposing the opening up of APIs to incumbent banks risks at once favouring start-ups, but also placing all financial services providers at a disadvantage compared to larger tech giants, who sit on top of large datasets they are not required to share. Similar situations may emerge in other sectors, such as energy.
- **One specific set of cases in which data-sharing is increasingly becoming an issue for policy-makers is complex, layered value chains in which Information Technology, and in particular AI and IoT, are becoming increasingly pervasive.** In those cases, the choice has to be made as to who will own the data that will be produced by the system good.

Acknowledgment of the principle of 'shared value creation' mentioned in the 2018 Communication "Towards a Common European Data Space" represents a first step, but its legal and economic consequences are still to be specified. For example, in agriculture the policy solution that seems to be prevailing is the attribution of data ownership rights to farmers; this, given the expected lower bargaining power of farmers vis-à-vis players at higher layers of the value chain, may lead to a more balanced distribution of the value created by high-tech (precision) agriculture.

- At the sectoral level, the European Commission is starting to promote **data-sharing arrangements as voluntary platforms** aimed at solving collective action problems and achieving economies of scale for the whole industry.[95] In this case, the Commission (or another institution, e.g. a sectoral agency) could act as orchestrator of so-called "industrial data spaces".[96] This could happen in sectors such as healthcare, for example through sharing of anonymized data and data on clinical trials; on energy, to enable a smarter managing of flows on the grid; on industrial B2B platforms e.g. to optimise logistics and facilitate coordination between large and smaller firms.
- Finally, and almost inevitably, there is a **need for skills and competences in data science and IT**, areas in which Europe seems to be unable to produce the needed talent, as well as unwilling to attract it from non-EU countries. This aspect will be discussed more in detail in section 5.4 below.

5.4 Research, innovation, education and society: Towards a new strategy?

As explained above, the EU strategy on AI is composed of two main pillars: the definition of ethical guidelines, and the promotion of European competitiveness in the field or Artificial Intelligence. However, **the rather optimistic tone with which the European Commission announced in April 2018 that Europe is very**

[95] For Intelligent Transport System, EU legislation already contemplates some data sharing obligations, See for instance Commission Regulation 2017/1926. The EP has called on the Commission to publish a legislative proposal on access to in-vehicle data (EP 2018a). With the release of the Third Mobility Package, the Commission announces a Recommendation that among other things will deal with "a data governance framework that enables data sharing, in line with the initiatives of the 2018 Data Package, and with data protection and privacy legislation."

[96] See for Fraunhofer's White Paper on Industrial Data Spaces and the corresponding reference architecture, https://www.fraunhofer.de/en/research/lighthouse-projects-fraunhofer-initiatives/industrial-data-space.html

well positioned to compete globally in AI is probably an overstatement. Both the United States and China dwarf the EU when it comes to research and development (R&D) investment in AI and related technologies, as well as in terms of uptake of AI-enabled solutions in the market. A study by McKinsey found that Europe's investment in AI totalled around €2.4-3.2 billion in 2016, compared with €6.5-9.7 billion in Asia and €12.1-18.6 billion in North America.[97] The US and China are also ahead of the EU in terms of investment in 5G, IoT and High-Performance Computing (HPC), with a real race emerging on quantum computing and even more (especially for China), on quantum cryptography.

Looking at data from the past decade, Europe seems to be extremely well positioned in at least one dimension, namely research publications in the field of AI. But this seems to be changing too, as Chinese researchers produce increasing amounts of new papers every year and file a remarkable number of new patent applications in this domain. A recent report by Elsevier confirmed that **Europe remains the largest producer of AI-focused research, with 30% share of the publication output in AI**. However, Europe risks suffering from an AI brain drain, losing many of its academic talent in this area, mostly to the corporate sector in the United States. The authors of the report also expect that **China will surpass Europe within four years as the leading AI-research geography** (Elsevier 2018).

What then should Europe do? **The CEPS Task Force converged on the need for Europe to avoid trying to compete with the US, Japan, Korea and China on all fronts**. Given the size of investment required and the level of advancement of these other regions, particularly the US and China, Europe does not seem to have any chance of leading on all fronts. Rather, its strategy should be more targeted and selective.[98] In particular:

- **Europe can try to lead, or at least compete at arm's length, in specific sectors** such as manufacturing, healthcare, transportation and finance. In those sectors, it should seek to establish standards, create industrial policy strategies and work on all aspects of the value chain, from infrastructure to data, skills, and applications/services.
- **Europe should play catch-up, for strategic reasons, in other specific sectors**. These include cybersecurity and defence. Significant resources should be devoted to the creation of capacity and resilience in these

[97] See McKinsey (2017), 10 imperatives for Europe in the age of AI and automation.

[98] At one of the meetings, John Zysman (UC Berkeley Center for Human-Compatible AI), discussed alternative strategies with Task Force members, such as imitating from behind, chasing the hype, or the so-called "Blue Ocean". These would have important consequences in terms of research, innovation and industrial policy.

domains, as well as in R&D at the pan-European level, in view of a future in which sovereignty of defence technologies and data may become essential.

- **Europe will inevitably have to "chase the hype" in some sectors**, mostly B2C ones, in which the US and China dominate the scene with very well-established tech giants. EU policy should, of course, continue to shape the regulatory framework to which these non-EU products will need to conform, if they want to keep operating in the EU: and this includes the rules and principles that will be introduced, or better explained, in the forthcoming Ethics Guidelines.

This three-pronged strategy should be accompanied by the right choices in terms of overall governance, in addition to regulation; and by important initiatives in the fields of research and education, innovation, and investment. We explore all of them below.

5.4.1 *Can Europe become attractive for AI research and innovation?*

As mentioned earlier, Europe has demonstrated a remarkable leadership in the number of research publications in fields related to AI over the past decade. At the same time, the Old Continent is losing positions in global university rankings, and Brexit would deprive it of the most vibrant AI research and innovation environment among member states. Very often, European researchers are forced to move to the US in order to have the chance to pursue a top-level career in academia, and when they manage to create start-ups out of their research ideas, they end up selling these ideas mostly in the United States.

The better pay offered by US and Chinese tech giants that are particularly active in research (Microsoft, Google, IBM, Huawei) is often too tempting for talented researchers to allow them to stay in their universities for long. And even more often, once researchers have left academia, it becomes impossible for them to go back to a more academic environment. Furthermore, many academic environments in Europe are still too confined in silos and are reluctant to undertake truly inter-disciplinary research: the defence of one's own turf becomes stronger than the desire to combine different perspectives, and thereby cover all relevant aspects of AI. There is still very little coordination in research funding between the EU and the national level, and even if one focuses only on the EU level, the fragmentation of research and innovation funding is reportedly an obstacle to the scaling up of research efforts in Europe. And finally, most countries in Europe have shown little openness to talent coming from abroad. Researchers very often ring-fence their departments to shield them from unwanted competition from the outside world.

All this would need to change before Europe can become an attractive place for researchers in the future, especially in cutting-edge fields such as AI.

In the years to come, EU institutions and member states will have to capitalise on existing knowledge and initiatives to create a new, flourishing environment for AI research in Europe. Policy and ethics can also play a role in this respect: as will be discussed in more detail in section 5.5 below, **if Europe clearly took the leadership on "AI for good", researchers with a strong motivation to develop AI-enabled solutions that address societal challenges would look more favourably at Europe**, especially if they could find a suitable research environment, a well-shaped policy context and enticing procurement and innovation markets.

In particular, besides the policies on data mentioned at the beginning of this section, the following actions appear to be important:

- **The issue of digital skills needs to be addressed, both for education and for research**. This is far from easy, for at least two reasons. First, the type of skills that will be needed in the future is by definition uncertain, given the fast-changing, often unpredictable direction of innovation in AI and related technologies. In particular, the emphasis placed until recently on STEM seems to be less justified, or at least not sufficient, for the next decade and beyond. Similarly, coding skills may become an inevitable addition to school curricula, but they are also unlikely to prepare society for the coming developments in AI. In any event, most of the coding, in the future, will be done by machines. Thus, it will be so-called "complementary, soft skills" that will matter the most: from humanities and social sciences to entrepreneurial skills, the development of our future workforce will have to focus on what makes humans really different from machines, and then nurture the human-machine interface. Second, education typically falls outside EU competences, and largely sits with national governments, irrespective of any attempt to achieve coordination and convergence from Brussels. It is only in the so-called Knowledge and Innovation Communities (KICs), managed by the European Commission's DG EAC (Directorate-General for Education, Youth, Sport and Culture) and in the European Social Fund, that the EU has managed to include education in large EU-level initiatives. It is therefore unlikely that action at the EU level, if not undertaken in a highly innovative way, will be able to fully promote the skills needed for the AI age. However, multi-level initiatives could be spurred through "agencification" – for example through the creation of a European Labour Authority, or the launch of a Mission on the Digital Transformation of Industry and Society, run by an agency (see below, section 5.4.3).
- **Universities should work on inter-disciplinary curricula**, which bridge computer science with natural sciences and social sciences, as often done

by universities outside the EU (e.g. in the US, many universities are now offering Masters degrees in inter-disciplinary data science). This may also entail the recruitment of hybrid profiles, with researchers who have accumulated experience in the private sector or in government, and in more than one field of studies. Creating diverse research and innovation teams with complementary, inter-disciplinary skills should not be a self-defeating choice for universities as it often is today, due to the silo approach that tends to favour homogeneous teams with publications in specific fields and in specific journals. The ability to experiment by mixing and matching academic backgrounds can only strengthen the human component of research and the value of its complementarity with AI and other data science tools. Similarly, competences in ethics and philosophy are poised to become increasingly essential in many fields of research, and some fundamental notions of these disciplines should be taught even in computer science courses.

- **Basic research should continue to be heavily funded in Europe**. Europe is currently leading in terms of public R&D funding of research (RISE, 2017), but severely lags behind private expenditure in R&D. The original goal (both in the Lisbon and in the Europe 2020 strategies) to achieve a combined level of R&D investment of more than 3% of GDP was never attained. Leading countries like Switzerland, Korea and Israel are way above the European average level of R&D expenditure. That said, the quality and direction of funding are also essential, and in the case of Europe the key determinant of funding choices, beyond the excellence of applicants, should be the extent to which research has the potential to tackle pressing societal challenges. The funding of basic research has proven to be of the utmost importance also for future development of commercialised products (Mazzucato 2014); and should be coupled with efforts to strengthen Europe's already well-developed AI community, and its relationship with civil society. Moreover, EU institutions should develop, together with member states, attractive visa programmes for non-EU talent: this could be done through the European Research Council, in the case of basic and applied research; and through the European Innovation Council for promising young entrepreneurs and innovators.
- **Researchers should be given a clearer career path and a smarter set of evaluation criteria**. The creation of spinoffs from university labs, as well as other entrepreneurial initiatives by researchers, should be encouraged as a sign of success, not of betrayal that distances the researcher from a university career. The evaluation of researchers and research teams should not be based only or predominantly on indicators such as the number of patents filed. And universities should be given significantly

greater funds in domains such as AI, so that they can compete with the private sector by offering an attractive mix of (greater) intellectual freedom and decent (even if slightly lower) salaries.

- **Research funding should focus on all aspects of the ecosystem, and explore more sustainable, human-centric, privacy-compatible ways of doing AI**. In particular, HPC (h**igh-performance computing)** funding should continue and be increased, with a view to developing alternative forms of computing power (e.g. neuromorphic chips), and advances in quantum computing (in which, however, the gap with private US and Chinese companies appears huge). New technological paradigms such as Edge computing and Fog computing should be included in public-private research and innovation actions, as they bear the potential to bridge cloud computing, distributed computing, blockchain and smart contracts, the IoT and data protection needs.[99]

All these initiatives, as mentioned above, require more than an evolution: in terms of its regulatory framework, Europe does not seem to need a drastic change, but when it comes to research, innovation and investment policy the need for discontinuity appears more evident. In particular, there is a need to build bridges "horizontally", i.e. between disciplines and through optimal portfolio management of various research paths; and "vertically", between the EU and Member States, between layers of the new stack; and between research and innovation, through knowledge transfer, future-proof policy-making, and ultimately sustainable innovation. Taken together, all these needs call for the adoption of ambitious, new path-breaking initiatives. These could take various complementary forms: a "CERN for AI"; an "AI-rbus", possibly dedicated to specific sectors such as healthcare, transportation and manufacturing; or a "Mission" in Horizon Europe, either generic or dedicated to the EU's most strategic sectors, e.g. a Mission on Healthcare, or a mission on Dementia, with a very strong AI and robotics component. These alternatives, which can also be seen as complementary, are explored in more detail below.

5.4.2 A "CERN for AI"

AI-related research and innovation in Europe needs talent, scale and direction. This calls for the creation of a catalyst, i.e. an institution or cluster of institutions that leads European development in this field. Concerning research, a set of initiatives already exists. For example, the Confederation of Labs of Artificial Intelligence Research (CLAIRE) and the proposed European Lab for Learning

[99] For a non-technical explanation and comparison of edge and fog computing, see https://medium.com/@thinkwik/how-edge-and-fog-computing-are-taking-over-traditional-cloud-computing-b26b7276f1ce.

and Intelligent Systems (ELLIS) are attempts to create networks of academic excellence in AI research. Both initiatives envisage distributed governance: for example, CLAIRE was launched by more than 600 of Europe's top AI researchers, calling for a substantial coordinated push for placing Europe at the top globally in AI-based research and innovation. The initiative has since then grown to more than 2,000, including 1,300 scientists and AI experts in academia and 500 experts from industry. Its structure has been defined as a "CERN for AI", due to its emphasis on national research laboratories connected with a strong central hub providing both the overarching infrastructure and a launch pad for mission-driven projects.[100] CLAIRE adopted a very broad focus on AI, rather than concentrating only on machine learning: the emphasis on the whole technology stack is extremely promising for Europe, and is based on the idea that "China and the USA are pushing two extreme and flawed models for AI", too data-dependent and almost exclusively focused on machine learning. This creates an excellent chance for Europe to define a middle way, balancing the interests of individuals, society and industry.

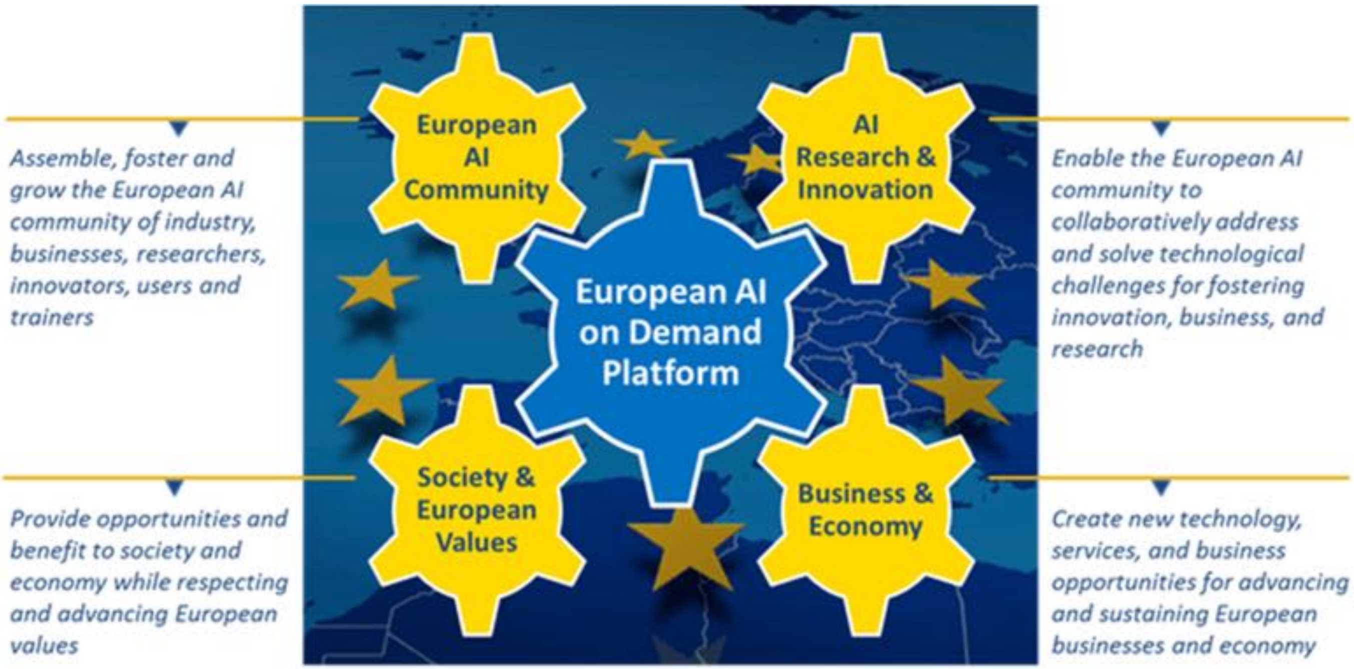
A number of activities in that direction have already been started: The "HumanE AI" proposal for an EU flagship project on AI has just progressed to the second phase, together with two other proposals in the AI context: robotics and language technology.[101] At the same time, another large project called AI4EU (which responded to a specific Horizon 2020 call) started in January 2019, with the goal to build a comprehensive European AI-on-demand platform to lower barriers to innovation, to boost technology transfer and catalyse the growth of start-ups and SMEs in all sectors through open calls and other actions; to act as a broker, developer and one-stop shop providing and showcasing services, expertise, algorithms, software frameworks, development tools, components, modules, data, computing resources, prototyping functions and access to funding; to offer training to enable different user communities (engineers, civic leaders, etc.) to obtain skills and certifications; and to establish a world reference, built upon and interoperable with existing AI and data components and other dedicated platforms. As diagrammed in Figure 10, AI4EU will mobilize the whole European AI ecosystem in the 27 member countries, including researchers, innovators and related talents, and launching pilots and research, together with an Ethical Observatory.

In section 5.4.3 below, we outline an ambitious proposal to scale up and align CLAIRE, HumanE and AI4EU through the establishment of a "Mission IT", nested in Europe's 2030 Agenda.

[100] Philip Slusallek, Scientific Director of the DFKI, presentation at CEPS Task Force.

[101] https://claire-ai.org/wp-content/uploads/2018/09/CLAIRE-Vision-Document-2-2.pdf

Figure 10. The AI4EU project – main goals



Assemble, foster and grow the European AI community of industry, businesses, researchers, innovators, users and trainers

Enable the European AI community to collaboratively address and solve technological challenges for fostering innovation, business, and research

Provide opportunities and benefit to society and economy while respecting and advancing European values

Create new technology, services, and business opportunities for advancing and sustaining European businesses and economy

Source: AI4EU (https://c1.assets-cdn.io/event/3394/assets/8459717489-92cef1529f.pdf).

5.4.3 A "Mission IT" on the digital transformation of industry

Regardless of whether research and industrial policy coordination is achieved at the EU level, there will still be important gaps in terms of how to achieve stronger scale and coordination in education and skills, as well as in innovation policy. This problem is not exclusive to AI: in many fields, the EU has struggled to achieve efficient innovation policy coordination, as well as an effective link between innovation and the whole policy process. There is also a need for an improved framework for private investment and enhanced cooperation between European companies to reach the necessary scale to compete with other global players. This is why in the future Horizon Europe programme, the European Commission is considering the launch of more structured "moonshots", or "moon landing" projects, in which the research, education, innovation and industrial policy components related to a single specific problem nested in societal challenges (mostly, the SDGs) are tackled through the formulation of specific targets and goals; and the appointment of an agency or portfolio manager with ample discretion on how to shape the mission. The five proposed missions presented in October 2018 to the Council were related to quantum computing, cures for childhood cancers, the elimination of plastic waste in rivers and seas, the creation of the first carbon-neutral cities with clean air and restoring soil health.

Could AI benefit from a specific "mission" in Horizon Europe? Probably not, since there is no single, easily specified set of milestones or achievements that can be easily associated with AI development. That said, AI can be essential to almost any mission that may be set by the European Commission (e.g. fighting dementia or optimizing transportation); and AI could feature as an essential component of a future mission, in particular a mission on the digital transformation of industry and society, with a strong emphasis on reskilling of the workforce and the analysis of the societal implications of the increasingly pervasive new technological stack. Milestones could then be expressed more easily (e.g. "reskilling 50% of the EU workforce by 2024", creating an immediate impact on European citizens in terms of visibility of the EU and its relevance, and added value. **The proposed mission would then have to be broader than AI (and also broader than quantum computing), embracing the whole technology stack, its ethical, societal and environmental impacts, and the possible consequences for education and skills. We call it provisionally "Mission IT".**

As part of a new trend, Mission IT would feature an embedded accountability dimension. Accountability refers in particular to the targets and milestones defined for the specific mission. Accountability, at the same time, requires constant feedback and evaluation, i.e. some form of performance measurement based on credible indicators. A system that allows for successful mission-oriented policy must then be "intelligent", in the sense that it should be able to learn and adapt over time, reflecting feedback on the changing conditions in the external context, as well as data on whether existing actions are producing the desired effect. Intelligence, here, thus means also that the system should be adaptive and flexible. In addition, an effective, adaptive and flexible system should also make room for experimentation.

Based on the second Memorandum of the "ESIR group", which provides advice related to the economic and social impact of research and innovation policy to the European Commission, missions should be asked to follow a cycle of road mapping, consultation, planning, experimentation, monitoring, evaluation, learning and feedback into the road mapping exercise.[102] This should be a constant cycle, which spins as fast as the mission allows, and should be fed by as many researchers and entrepreneurs as possible. More specifically, missions should lead to extensive experimentation of possible solutions to the problem identified. This responds both to a logic of risk management (different solutions, with different levels of risk and reward, should be tried at the same time), and to a logic of more inclusive innovation policy (the whole EU community or researchers and innovators should potentially be involved in trying to find a solution to the problem). Experimentation could follow two tracks: i) experimenting with new technologies/business models/delivery modes and blending funding instruments and schemes to run experiments;[103] and ii) experimenting with policy solutions, by engaging in experimental policy-making and inspiring legislative proposals that would remove obstacles to promising solutions. Policy experiments could include instruments such as randomized controlled trials, rapid prototyping, landscaping, ideation sprints, instant focus groups, scenario testing, virtual and actual sandboxes and randomized controlled trials are of utmost importance for the future of innovation-friendly policy-making, together with algorithmic approaches to regulation (Yeung 2017). The overall idea is to generate experience and data, which will later enable counterfactual evaluation of the prospective, possible

[102] See ESIR Memorandum (2018), Implementing EU Missions, at https://ec.europa.eu/info/sites/info/files/ki0618012enn.pdf

[103] This could happen on a "prize" basis, or on a more top-down selection of possible paths (e.g. technology roadmap), or both. For example, the replacement of general practitioners with online, constantly available bots could be subject to experimentation with a sample of patients, carefully selected. In terms of instruments, the expectation is that missions will be able to tap into various sources of funding, including research funds, EIC funds, EIB, InvestEU, structural and cohesion funds, national funds made available on a voluntary basis by member states and even non-EU countries (in the spirit of "Open to the World"), and private funds (partnerships). The ability to blend different forms of funding would be considered as essential to the skills and activity of the mission.

impact of the new solution. The input to policy-making could take the form of a "wish list" that would be submitted to DG RTD and later to the Secretary General for inclusion in the Commission yearly work programme.

A key choice would then be the appointment of a very capable, charismatic leader for the Mission IT. Van Atta (2007) describes in this way the figure of a manager of a mission-oriented agency like DARPA: "The DARPA program manager is, in fact, the key. [S]he is the technical champion who conceives and owns the program. [S]he is not told what to do, though [s]he does have to have approval from his/her office director, and from the DARPA Director. Once [s]he starts that program, it is his/hers, and [s]he makes it happen, and [s]he has to make the choices involved in that. So, in essence, they are risk-taking, idea-driven entrepreneurs heading up their own practice."[104] The key issue is to make missions attractive for world class talent, who will be interested to become mission leaders. Leaders should help to develop human resources, talents and skills.

In terms of governance, **it is still unclear whether future missions under Horizon Europe will be run by stand-alone agencies or bodies, or set up as independent entities (like JTIs or KICs). Transparency and accountability reasons would lead to the conclusion that this should be the case.** In the case of a "Mission IT", an agency or a portfolio manager would:

- Coordinate the research dimension on all layers of the technology stack, including HPC, connectivity and 5G, data flows and platforms (including blockchain/DLTs), AI and IoT;
- Explore alternative avenues for research, e.g. less data-hungry AI techniques, advanced research on natural language processing, alternative chips (including a European chip[105]), interaction between AI and blockchain, cryptographic solutions for privacy-preserving algorithms, debiasing techniques, etc.
- Work to stimulate knowledge transfer between advanced research centres (including the possible distributed structure of the CERN for AI) and entrepreneurs and SMEs in all member states, like Knowledge and Innovation Communities (KICs) have done in the past years;

[104] Van Atta, R. H., 2007. *Testimony before the Hearing on "Establishing the Advanced Research Projects Agency-Energy (ARPA-E),"* Subcommittee on Energy and Environment, Committee on Science and Technology, United States House of Representatives, Washington, D.C. Please note that we have rendered his quote gender-neutral

[105] https://www.reuters.com/article/us-europe-tech/european-chip-industry-seeks-more-eu-help-to-extend-revival-idUSKBN1JN1SD

- Producing standards (in cooperation with ETSI), where possible replicating international (e.g. IEEE standards), adapted to the EU approach to AI;
- Integrate an education dimension, post-graduate courses and PhDs to be integrated in highly vibrant, interactive research teams;
- Monitor the evolution of technology and provide ethical advice, by incorporating the AI HLG and reaching out constantly with the AI Alliance. This could then become a way to keep the "live" sections of the Ethics Guidelines constantly updated.
- Create a bridge between the EU and the member state levels of research and innovation funding, and policy, by creating synergies and helping locate research centres and teams where member states provide co-funding and express specific strategic interests.

Finally, and importantly, the **Mission IT would be complementary to other missions**, e.g. a mission on reducing the burden of dementia would profit enormously from AI developments, as well as breakthroughs in computing capacity, including neuromorphic chips.[106] Mission IT could become instrumental in optimising future traffic flows, helping with big data and AI to predict future epidemics and pandemics outbreaks (Fiorillo et al. 2018).

5.5 The Commission's Coordinated Plan – a brief analysis

The European Commission, in cooperation with EU member states, adopted a Coordinated Plan in December 2018, with the aim to "maximise the impact of investments at EU and national levels, encourage synergies and cooperation across the EU, exchange best practices and collectively define the way forward to ensure that the EU as a whole can compete globally".[107] This is a very welcome development, especially given the level of fragmentation that national AI strategies had exhibited until the launch of the Communication on AI in April 2018. Today, Europe seems to have taken the AI opportunity seriously, and is taking steps to boost investment in research and innovation, to make AI a cornerstone of future EU growth.

The first, resounding commitment in the Coordinated Plan is related to investment. The Commission plans to scale up investment to reach the target of €20 billion per year over the next decade; and increased investment in AI under Horizon 2020 to €1.5 billion in the period 2018-2020, with a view to reaching €20

[106] See a non-technical definition of neuromorphic chips at MIT Technology Review, https://www.technologyreview.com/s/526506/neuromorphic-chips/

[107] https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence.

billion for the period 2018-2020 if member states and the private sector make similar efforts. The Commission also proposed to invest in AI at least €1 billion per year from Horizon Europe and the Digital Europe programmes in the next MFF.

Importantly, without referring to "missions" as this report does, the Coordinated Plan mentions the need for a new research and innovation partnership on AI, fostering collaboration between academia and industry. The Commission also announced that it will bring companies and research organisations together to develop such a common strategic research agenda. The agenda will build on existing partnerships in robotics and big data, representing an investment of €4.4 billion, of which the majority comes from the industry. This proposal is coupled with announced tighter networks of European AI research excellence centres (which echoes the so-called "CERN for AI" described above); support for large-scale pilots and experiments in areas such as smart farming, smart cities and connected and autonomous vehicles as part of the implementation of the Digitising European Industry strategy; and the development of several large-scale reference test sites, using up to €1.5 billion from the AI strand of the Digital Europe programme.

Other announced initiatives include the creation of "common European data spaces" in areas such as manufacturing or energy, which are expected to become a "major asset for European innovators and businesses". These data spaces are defined as instruments that aggregate data both for public sector and for B2B across Europe and make them available to "train AI on a scale that will enable the development of new products and services".[108] This will require the rapid development and adoption of European rules such as interoperability requirements and standards, and the contribution of high-value data sets by member states and the Commission (e.g. earth observation data and information from the Copernicus programme). In 2020 the Commission will support via Horizon 2020 the development of a common database of health images, which will be anonymized and will be dedicated to the most common forms of cancer, using AI to improve diagnosis and treatment. Other initiatives include joint procurement of AI solutions; and initiatives on computing capacity, based in particular on the European High-Performance Computing Initiative (EuroHPC), the partnership with member states and industry on microelectronic components and systems (ECSEL) as well as the European Processor Initiative.

The CEPS Task Force welcomes the Coordinated Plan and its high level of ambition. The suggestions and recommendations included in this report, both on the draft ethical guidelines and on the policy and investment side, are aimed at contributing to good governance and more granularity in the implementation

[108] http://europa.eu/rapid/press-release_MEMO-18-6690_en.htm.

of these action items. This includes the "mission IT" idea, which may lead to more effective coordination in drawing links between the High-Performance Computing side, the AI-related initiatives (both research and ethics) and other elements of the ecosystem, including IoT, data spaces, 5G connectivity, etc. It would also facilitate the bridging of education, research, innovation and policy under the same broad umbrella initiative, with key milestones and targets, effective stakeholder involvement and inclusive efforts to create a European IT ecosystem.

The only big element that really seems to be missing in the puzzle of actions portrayed by the Coordinated Plan is the link between AI and the SDGs. This is regrettable, since – as has been argued in various sections of this report – such a link could become the most distinguishing feature of the EU's approach to AI and would establish Europe in a prominent position on AI policy at the global level, including vis-à-vis those countries with which the EU has consolidated trade agreements, or is about to engage in deeper trade negotiations. The next section comments on Europe as a possible leader in this field.

5.6 Can Europe be the champion of "AI for Good"?

The previous sections have described possible initiatives that bear the potential to realise Europe's ambition to lead, or at least compete, in certain aspects of trustworthy AI, and fully capitalize on its current research excellence in specific AI fields. But **the domain in which Europe could really fill a gap, and try to lead the rest of the world, is the alignment between AI (and all IT) with economic, social and environmental goals, such as the SDGs**. As Europe has already committed in 2016 to mainstreaming SDGs into every aspect of EU policy (as observed above), the time is ripe to practice what EU leaders have preached and launch a substantial effort in the mapping of how all digital technologies can help Europe and the world achieve the ambitious 2030 goals.[109] This effort would also deeply resonate with the EU's external action: the Global Strategy on Foreign and Security Policy for the European Union sets out the strategic direction for the EU's external action and identifies clear links to the 2030 Agenda. It emphasises the importance of a comprehensive approach in the EU's external actions and the need for an integrated EU approach to increase the EU's impact in responding to and preventing violent conflicts and crises as well as of improving coherence between the EU and its member states. The new

[109] See the European Commission Reflection Paper, Toward a Sustainable Europe by 2030, at https://ec.europa.eu/commission/sites/beta-political/files/factsheets_sustainable_europe_012019_v3.pdf.

European Consensus on Development put forward a shared vision and framework for action for all EU Institutions and all member states, framed around the five key themes of the 2030 Agenda: people, planet, prosperity, peace and partnership. It places particular emphasis on cross-cutting drivers of development, such as gender equality, youth, sustainable energy and climate action, investment, migration and mobility, and seeks to mobilise all means of implementation: aid, investments and domestic resources, supported by sound policies.

Despite this bold commitment, as mentioned in section 4 above, there has been no real mainstreaming of the SDGs so far (with the exception of a timid attempt in the discussion on the MFF); and what is more important for the purpose of this report, there has been no real mapping of what digital technology can do for SDGs, and in particular of what AI can do to achieve what we termed "level 3-compliant" AI. The link between AI and the SDGs features, although not prominently, in the JRC Flagship report on AI, but other than this, no other EU document has attempted to approach AI as a means to 2030 ends, rather than as an end in and of itself. At the global level, the "AI for good" initiative remains rather isolated, and a gap exists, which a strong effort at the EU level could fill. This would be potentially beneficial, in terms of competitiveness, sustainability and also as a new narrative for the EU.

The ideal governance arrangement for making Europe a champion of "AI for good" and "AI for the SDGs" is the one outlined in the previous section, for the following reasons. First, future missions will be explicitly nested in the SDGs, as mentioned in the two ESIR memoranda, in the Mazzucato report, and throughout the Horizon Europe proposal.[110] Second, the need to achieve multi-level coordination and a coordinated, inclusive approach to research, education, innovation and industrial policy finds a good compromise in the creation of an entity external to the European Commission, able to guarantee a fast rotation of guest researchers from all over Europe and beyond. Academic literature (Fuchs 2010; Sen 2017; Azoulay et al. 2018) has found, based on analyses of mission-oriented agencies such as ARPA (Advanced Research Projects Agency) in the US, that it is possible to organize research and innovation efficiently around technology-related missions on the basis of a set of overarching goals. Moreover, a mission-oriented approach can be optimal for technological areas where technology exists but is relatively unexplored and has great potential for

[110] See the Mazzucato Report, Mission-Oriented Research & Innovation in the European Union, at https://ec.europa.eu/info/sites/info/files/mazzucato_report_2018.pdf; and the Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination, COM/2018/435 final.

improvement. Finally, a mission-oriented approach can help solve friction in markets for ideas and technologies in sectors where the path from idea to impact is extraordinarily difficult (such as in energy because of numerous obstacles such as the large amount of capital required for demonstration and scale-up, strong infrastructure inertia, etc.).

Failure to recognise and publicly promote the role of AI and its related technologies for a more sustainable future society would represent an enormous missed opportunity for Europe. Of course, AI should not be the only means to achieving the SDGs: there should be ad-hoc social and environmental policies that cater to the needs of society. That said, until now the debate has mostly focused on avoiding potential harm, rather than proactively promoting AI's contribution to a more prosperous society. A perfect example of this mismatch is the absence of SDGs in the Draft Ethics Guidelines, as well as in the Coordinated Plan adopted in December 2018 by the European Commission.

PART III.
THE WAY FORWARD

This final part of the report draws on the analysis and findings of the previous sections and presents the main recommendations of the CEPS Task Force on AI for the future EU Strategy. These are intended as constructive input into the ongoing process taking place now between January and March 2019, with the aims of refining the EU Draft Ethics Guidelines on Artificial Intelligence and of defining the main policy and investment recommendations for the future AI strategy before mid-2019. This final contribution is divided into two parts: section 6.1 outlines our main general recommendations on the EU Draft Ethics Guidelines, and section 6.2 contains our recommendations for future EU policy and investment priorities in this field.

6. MAIN RECOMMENDATIONS

6.1 Recommendations on the Draft Ethics Guidelines

The main assumption behind the current EU strategy on AI is that "Europe can lead", which in turn requires three separate, but complementary commitments: i) to increase investment up to a level that matches Europe's economic weight; ii) to leave no one behind, in particular when it comes to education and ensuring a smooth transition towards the AI age on the workplace; and iii) to base new technologies on "values". In December 2018, Draft Ethics Guidelines produced by the High-Level Expert Group on AI were published. We present below the main recommendations of the CEPS Task Force on the future guidelines. Such recommendations apply under the assumption that the Ethics Guidelines will remain a non-binding document.

1. **It is important that the European Commission, backed by the High Level Expert Group on Artificial Intelligence, does not limit itself to adopting simply another list of principles. Instead, it should strive to develop a list that truly represents the EU's approach to AI and offers concrete guidance** to stakeholders. This guidance should identify which applications or business models are potentially problematic and which ones should be altogether prohibited as they are incompatible with EU core values and legislation. Our report has explored several fundamental questions: What are the concrete implications of the forthcoming Guidelines for AI development, both overall and for different use cases? How could developers ensure that they are adhering to the Guidelines? Will they be translated into enforceable policies and if so, who will enforce those policies and how?
2. **The future EU Ethics Guidelines on AI should be essentially addressed to the "supply side"**, including AI developers, vendors, and distributors; and also to organisations using or deploying AI; and to public administrations, which should decide whether, and how, to use AI in their daily activities, and how to procure AI products in a way that is aligned with EU values and existing legislation. **The Guidelines should not contain separate sections for different types of actors**. Rather, the Guidelines should include a definition of values and principles that developers, vendors, distributors of AI, as well as organisations using or

deploying AI can adopt as reference when placing AI systems on the market (whether B2B or B2C) or using AI systems in-house to improve their operations.

3. **This, of course, does not mean that initiatives should not be taken on the "demand side"**. Those on the receiving end should be able to discern when the AI system that is being used does not comply with fundamental principles, and those cases in which this would allow them to seek redress since the system in use violated their rights as protected by the EU legal system (irrespective of whether an ad-hoc system is in place for users to seek redress).
4. **The forthcoming EU Ethics Guidelines should include four main sections, spelling out the following**: i) EU values and principles of responsible, accountable and sustainable AI; ii) AI applications and use cases that appear to be problematic from the standpoint of the application of the identified principles, as well as use cases in which there is no specific ethical problem raised by the AI system (but of course the system has to be compliant with EU legislation anyway); iii) guidance on possible measures that could be adopted by AI developers, vendors, distributors and organisations deploying AI to check that their AI-enabled applications and systems are aligned with the ethical guidelines; and iv) users' rights and possible enforcement measures and channels for redress in case of infringement of those principles outlined in the guidelines that are mandatory based on EU law.
5. **The Ethics Guidelines should be available as an online, living document**, in order to reflect developments in AI applications and technologies, the development of best practices and the identification of potential issues. This will also require constant support for this activity, which could be provided by the Commission, by the High-Level Expert Group or by future institutions or agencies such as the one involved in the "Mission IT" proposed in this report (see Recommendation no. 43 below); or the "Ethics hub" currently being discussed by the EU High Level Expert Group on AI.
6. **In the Ethics Guidelines, the European Commission should identify core values and principles at three different levels:**
   - ***Level 1: Fundamental principles (Lawful AI).*** One first definition of those principles is that, being rooted in the EU core values, they are in any event *mandatory* for AI developers, vendors, distributors and organisations deploying AI. As a preliminary list, these principles should include: the "non-maleficence principle" ("do no harm"); protection of human integrity, security and privacy; principle of respecting human dignity, including the right not to be discriminated against; and protection of agency, freedom and the democratic process.

- *Level 2: Good practices in AI development* **(*Responsible AI*).** This set would include some of the principles that have been agreed upon by AI developers in documents such as the Asilomar principles. As a preliminary list, level 2 should include the following principles: principle of complementarity with humans ("human-centric AI"); responsible and agile governance; monitoring, control and feedback practices; principle of AI transparency and explainability.
- *Level 3*: ***Principles of sustainable AI (Sustainable AI).*** This set of principles marks the difference between AI that is allowed to circulate and be implemented in the EU space, and AI that is fully aligned with EU medium-term policy goals (2030 Agenda). They include, as a preliminary list, the following: the "do good" or benevolence principle; AI that features limited or zero carbon footprint (SDGs 7 and 13); AI that promotes inclusive growth, full and productive employment, and decent work for all (SDG 8); AI that promotes quality education (SDG 4); AI that fosters women empowerment (SDG 5); and AI that contributes to industry, innovation and infrastructure (SDG 9). In some circumstances, the EU may require adherence to this most ambitious set of principles: this could be the case, for example, of AI used by public administrations; or publicly procured AI.

7. **The second part of the Ethics Guidelines could include a list of use cases** that are considered to be representative, borderline, or problematic by the EU High-Level Expert Group. Such use cases could be divided into two sections: i) prohibited use cases, such as Lethal Autonomous Weapons or the full delegation of certain "life or death" decisions to algorithms; and ii) "problematic" or "borderline" use cases, to be collected i.a. through the AI alliance, including predictive policing, social credit scoring, non-targeted use of facial and body recognition, etc. This second group could then become a candidate for future pilot tests, or regulatory sandboxes.
8. **The third section of the Ethics Guidelines could point at possible arrangements that would significantly mitigate the risks posed by a specific use of AI**. There is reason to believe that the first strongholds against bias creeping in corporate algorithms come with embedding good practice into companies' daily risk management activities. Guidance could also be offered to developers, vendors and SMEs as to when techniques such as unsupervised deep learning and reinforcement learning are appropriate, and what consequences occur in terms of explainability of AI. Finally, guidance could also be offered with respect to making algorithms GDPR-proof by using specific cryptographic techniques.
9. **The fourth and last section of the Ethics Guidelines could include guidance for the end users**, integrated with existing and new tools to

educate them about their rights and what to expect when dealing with AI-enabled systems. This section, too, could be usefully fed by the contributions from end users and other stakeholders (e.g. SMEs, developers) through the AI Alliance.

10. Based on the work of the CEPS Task Force, it is argued that **the current list of core values and principles in the Draft Ethics Guidelines lacks hierarchy between principles and values that should always be adhered to as they reflect constitutional values and principles of the EU; and those that correspond to good practices in AI development, or are aligned with EU medium-term policy goals**. This differentiation would add considerable value to the Guidelines, which otherwise would end up re-proposing the same structure of the rather general list of principles listed by the EGE group, and also by many other governmental and non-governmental documents over the past years. It would also, at least partly, remedy the fact that the current assessment list appears too long and is in strong need of a reformulation.

11. **The principle of non-maleficence is currently defined in the Draft Ethics Guidelines in a way that is hardly actionable for policy-makers**. More clarity would be needed for all those borderline cases in which an AI system ends up in a "life or death" or even more in a "death or death" type of dilemma (as in the trolley problem). The same principle ("do no harm") – although it is environmentally-friendly, it approaches sustainability only from an environmental dimension, with no reference to social and economic aspects. Moreover, the inclusion of environmental sustainability under the non-maleficence principle does not necessarily imply that these principles will be fully enforced in practice.

12. **The Draft Ethics Guidelines contain provisions that are difficult to translate into concrete policy.** If the principles (for example, those on human agency and self-determination**)** as they stand are effectively enforced, this would lead to massive over-regulation. Conversely, if they are not made enforceable, they might remain pure "wishful thinking" just as many other declarations of principles and values. The CEPS Task Force proposes to establish a hierarchy of principles that allow EU institutions to tailor their policy approach in a proportionate way, based on the specific use case, as well as on the organization or institution deploying AI systems.

13. **The idea that all AI systems should be gauged against fairness standards should be accompanied by extensive explanations** on key aspects of AI development and commercialization, such as what is acceptable bias and when does bias become unacceptable; what is a fair outcome taking into account the context and use case; and which corporate practices are deemed to appropriately tackle the problem of bias.

14. **Accountability mechanisms are briefly presented, mostly in the form of legal liability.** They should be expanded to cover all forms of accountability throughout the process of AI development. Also, the legal liability part, which is more developed, could be significantly improved. For example, the treatment of compensation of moral damages is absent, and is briefly replaced by a statement according to which "in a case of discrimination, however, an explanation and apology might be at least as important". Even more importantly, the availability of redress for individuals should be addressed more in detail and be linked to the principles of accountability and explainability.

15. **The Ethics Guidelines should not require AI to always be designed for all.** This provision appears to be disproportionate at best, and likely to impose unnecessary costs to all developers of professional AI systems, or even hamper AI design aimed at augmenting humans in specific professions. That said, there may be specific cases in which design for all could be required, as in the case of public services.

16. **The Draft Ethics Guidelines, while correctly mentioning both intentional and unintentional discrimination, do not explain what constitutes acceptable discrimination, and what forms of discrimination are unacceptable**. Clarifying those aspects is exactly what the Draft Ethics Guidelines should do to increase the level of certainty and trust in the way AI interacts with society and the economy.

17. **The future Ethics Guidelines should clarify in detail when personalisation services** through, e.g. recommender systems, search engines, navigation systems, virtual coaches and personal assistants **are to be considered disproportionate or excessively harmful to human autonomy**: currently, the exact definition of user well-being and the cases in which the system excessively nudges the user are not clear.

18. **The Draft Ethics Guidelines seem to imply that robust, trustworthy AI provides for the replicability of results**, since "lack of reproducibility can lead to unintended discrimination in AI decisions". However, it is not clear how far-reaching is this requirement; and to what extent it is being linked (as should be the case) with explainability. The same applies to the references to "accuracy" and to the need for a "fall-back plan": more detail will be needed in order to make these principles actionable. Given that trade-offs may emerge between, e.g. explainability and effectiveness of AI, guidance on how to strike the right balance in given use cases would be essential.

19. **The finalisation phase of the Ethics Guidelines should feature an in-depth discussion on how to deepen guidance and discussion on the "ten requirements" of trustworthy AI proposed by the HLEG, as well as on**

what happens if the requirements are not met. The current list of requirements appears difficult to implement, and the ten requirements appear to overlap in many respects. Once the Guidelines are finalised, there should be a mechanism for stakeholders to test them and eventually decide whether to adhere to them. The Guidelines could also usefully translate into a standardisation process at the EU level, as well as into a code of practice.

20. **The structure proposed for the Guidelines by the CEPS Task Force (see Recommendation no 5 above) offers advantages compared to the current draft.** These are mostly related to features such as scalability, flexibility, and modularity; to legal certainty; and also to the greater level of detail, easier translation into endorsement or enforcement mechanisms, and a strong link with sustainability. With a clear distinction between legal, responsible and sustainable AI principles, the High-Level Expert Group could also suggest a more tailor-made policy for specific use cases. For example, sustainable AI (which also incorporates, in our scheme, "Lawful AI" and "Responsible AI") could be mandated only for AI implemented or procured by public administrations; or in specific sectors such as healthcare services; whereas for most other uses, "legal AI" would suffice. Clarifying which level of care is expected for different use cases would also improve clarity in shaping the legal framework for liability.
21. **While the "AI as the new GDPR" vision is interesting, it is not necessarily compelling.** After all, it is too early to conclude that the GDPR has become a global standard, and it is also too early to reach conclusions on the impact of GDPR on Europe's competitiveness in the digital sphere. A proportionate approach is needed, aimed at steering AI towards the common good when needed, but also taking care not to overburden innovators and entrepreneurs with procedural requirements and compliance costs. Already providing guidance to AI developers, vendors and distributors on how to ensure that AI is compatible with the GDPR would be a major step forward. In this respect, rather than AI as the new GDPR, it would be important that AI and GDPR go hand in hand in Europe.
22. **The CEPS Task Force did not find sufficient grounds to suggest the adoption of public certification, or even mandatory standards on AI in Europe**. The main reasons are that a number of organisations, including IEEE and ISO, are developing their own standards; organisations or academic associations have developed principles (e.g. Partnership on AI, the Asilomar principles); and the market seems to be generating self-certification frameworks and packages, with companies that are developing step-by-step guidance and solutions for firms that adopt AI-enabled systems as well adopting their own principles for how they develop and use AI. Moreover, one could expect the forthcoming EU Ethics Guidelines

to represent an additional benchmark for corporations and intermediaries, who could then signal to their customers the alignment of their product with most or all of the principles included therein. Finally, it is still too early to anticipate with reasonable certainty how the AI market will develop over time, and in various domains and sectors: one possibility is that IT firms and consultancies will end up supplying most of the AI solutions in the form of pre-trained algorithms, for use by SMEs, which would lift the burden of value alignment of AI systems for SMEs.

23. **So-called "monitored self-regulation" is the preferred option for the CEPS Task Force**, when it comes to the promotion of the Ethics Guidelines. The European Commission should first evaluate the situation of the services offered today to EU citizens, and its likely evolution over time. Only if it emerges that those services are unlikely to comply with the Guidelines should the Commission consider additional policy actions. This does not mean, however, that the EU should not take any policy measures, for example, to clarify the issue of liability for damages caused by AI systems or to encourage data-sharing in specific sectors.

24. **The preferred solution at this stage is to require different levels of ethical alignment, depending on the use case: "level 3" alignment or "Sustainable AI" should be required only in very specific contexts, such as the AI-on-demand platform, public procurement of AI solutions for public administrations and the delivery of public services; and the use of AI solutions in research, innovation and investment policy**. This very ambitious set of ethical standards could also be required in specific sectors, such as healthcare, subject to a careful impact assessment. The European Commission should also promote the Ethics Guidelines by mobilizing the AI alliance, coordinating with member states and raising the awareness of the end users through examples and use cases that explain the importance of ethically aligned AI. The Commission should also monitor the market to ascertain whether the Guidelines are having the expected impact, and if it determines that the measures are insufficient to steer the market towards accountable and sustainable uses of AI, it should then consider alternative measures such as co-regulation.

25. **The Commission should foresee adequate room for experimentation**, in the form of "regulatory sandboxes" or randomized controlled trials, to ensure that AI-enabled solutions that potentially pose ethical concerns prove their value in terms of user protection before being admitted to the market. These measures, however, are not directly related to the Guidelines and accordingly are dealt with in the next section of this report.

6.2 Recommendations for future EU policy and investment priorities

The second deliverable of the EU High Level Expert Group on AI will be dedicated to the formulation of recommendations for future EU policy and investment priorities. This final section briefly looks at policy changes that will be needed in order to promote and implement the EU's approach to AI, starting from general considerations about the EU's Better Regulation agenda and then digging more deeply into the issue of reforming legislation and reflecting on governance issues.

26. **The CEPS Task Force did not find strong evidence that would favour a massive revision of existing horizontal regulations in fields such as product liability and machinery**. Accordingly, there seems to be a need for an evolution, rather than a revolution. However, it must be recognised that technology is also bringing important changes in the way governments approach regulation, and this will affect the way in which AI policy is likely to evolve over time.
27. **The EU's Better Regulation tools should be adjusted to reflect the specificities of AI.** This applies at all levels of the EU policy cycle. In particular:
    - The problem definition phase should be increasingly based on foresight and risk analysis.
    - The Commission should remain open to a new form of "innovation deals", in which AI developers challenge existing regulation by showing that they can achieve significant benefits and high levels of user protection with alternative business models than the ones on which the original regulation was based.
    - An experimental phase should be foreseen, in which new co-regulatory solutions and/or new business models are tested in a secure space such as a sandbox, before they are admitted to the market.
    - The Commission should analyse possible changes in the AI-related policy framework avoiding exclusive reliance on cost-benefit analysis and relying more on multi-criteria analysis: regulatory options should lead to a high level of protection of basic and fundamental rights, as well as additional policy objectives such as user empowerment and 2030 Agenda sustainability goals.
    - It is very important that the Commission respects the so-called "Treaty-based principle of proportionality", which dictates that any

proposed means of intervention be proportionate to the stated goals, and thus avoids being overly prescriptive or invasive.

- o Risk analysis should become way more deeply embedded in the regulatory practice of the European Commission, or of a dedicated new agency.
- o Policy frameworks should be adaptive and feature a strategy for data collection in order to enable monitoring and possible changes over time. In the case of co-regulatory or monitored self-regulatory schemes (including RegTech or SupTech), the private sector should cooperate with EU institutions in order to enable seamless monitoring of existing policy solutions.

28. **The liability regime for AI products and services should not be based on fault, but rather on relative strict liability, in particular for B2C use cases.** However, the Commission should clarify under what circumstances tortfeasors will be exempted from liability; when liability exposure will be mitigated by contributory negligence; whether there should be single entities (e.g. AI vendors, or producers, or vendors of AI-enabled system goods such as self-driving cars) responsible for compensating end users; and also what happens in case of damage caused by the interaction between AI systems (so-called "flash crashes").
29. **The design of a liability regime for AI inevitably boils down to a fundamental question: Should AI be treated, from a legal perspective, as an extension of the human being, or as a part thereof; as equivalent to a product or a service; as equivalent to an animal; equivalent to a slave; or tantamount to an employee, with legal personhood**? The current EU legal framework appears largely adequate, but it may need some clarification and interpretive guidance in order to avoid causing confusion and uncertainty among industry players.
30. **The future EU liability regime will also have to be designed in combination with a suitable insurance framework**. The future Ethics Guidelines could advise any company developing, embodying or selling AI in their systems to verify their financial ability to respond to potential liabilities that could arise from its use. If that is not possible, users should be required to abandon that use or to cover those risks with an insurance or an equivalent requirement. And if the insurance system ends up being too burdensome, especially for SMEs, a mandatory, subsidized insurance system should be foreseen, in order to combine the benefits of innovation with the certainty of compensation for end users.
31. **The Commission should avoid introducing new regulatory requirements that are additional and inconsistent with existing sectoral rules** on transparency, accountability and non-discrimination in regulated sectors

such as banking, insurance, healthcare, etc. One possibility would be to work in the direction of a "tech REFIT", i.e. an expansion of the ex-post evaluation methodology, aimed at introducing specific questions regarding the compatibility between existing legislation and the present and likely future developments of AI. Such tech REFIT could be added to the current better regulation guidelines, as well as to the work of the REFIT platform.

32. **It is of the utmost importance that governments adopt *open data* policies, by making large datasets available to the public, possibly in formats that are interoperable with existing machine-learning software**. So far, data held by government and data from publicly funded research are still largely unavailable for researchers, entrepreneurs and companies willing to engage in data-driven innovation.

33. **The free flow of data in the Single Market should be further promoted**, in line with the ambition of the Regulation adopted in November 2018. At the same time, the possible exceptions to the free flow, for example based on national security stances, should be narrowly interpreted to avoid disproportionate disruptions of data flows.

34. In order to reconcile data availability with the need for data protection, the **European Commission should fund research, innovation and standardisation in the domain of privacy "by design", as well as privacy-enhancing technologies**. In the case of large datasets, key technologies include cryptographic solutions that allow for the use of large datasets without infringing privacy laws, such as private practical computation, zero-knowledge proofs, homomorphic encryption.

35. **Allowing text and data mining for both research and commercial purposes would be very important for data-driven innovation in Europe**. Data mining is essential for the future competitiveness of the EU: if subject to ethical guidelines (so-called "Ethical Data Mining"), it could become an engine of data-driven innovation in Europe.

36. **Experimentation is key to keeping Europe relevant in the AI field**. However, the GDPR's data minimization principle and the need to have a clear purpose for obtaining user consent can limit the ability to experiment with innovative approaches, even when users have given explicit consent to access their data.

37. At this stage **it seems wise to avoid more strictly regulating access to data**, in particular outside the rather confined remit of competition law and refusals to deal by dominant companies. The recent Communication Towards a common European data space (EC 2018) argues that in general stakeholders do not favour a new data ownership type of right and indicate that the crucial question in B2B sharing is not so much about ownership, but about how access is organised. In any event, there seems to be a need for

more clarity on the legal framework for machine-generated data: the third Data Package adopted by the European Commission in April 2018 focuses on the review of the Public Service Information (PSI) Directive for data held by the public sector and on soft law for access to and preservation of scientific information; as to access and re-use of private sector machine-generated data in B2B relations, the Commission has defined a series of key principles that should be respected in contractual agreements in order to ensure fair and competitive markets, and separate key principles for data sharing in business-government relations.

38. **The issue of data sharing is becoming a reality, especially in complex, layered value chains in which Information Technology (IT), and in particular AI and IoT, are becoming increasingly pervasive.** At the sectoral level, the European Commission is starting to promote data-sharing arrangements as voluntary platforms aimed at solving collective action problems and achieving economies of scale for the whole industry. In this case, the Commission (or another institution, e.g. a sectoral agency; or an ad hoc body along the lines of the UK Open Data Institute) could act as orchestrator of so-called "industrial data spaces".

39. **Europe should avoid trying to compete with US, Japan, Korea and China on all fronts**. Rather, the strategy should be more targeted and selective. In particular:
    - Europe can try to lead, or at least compete at arm's length, in specific sectors such as manufacturing, healthcare, transportation and finance. In those areas, it should seek to establish standards, create industrial policy strategies, and work on all aspects of the value chain, from infrastructure to data, skills, and applications/services.
    - Europe should play catch-up, for strategic reasons, in other specific sectors, including cybersecurity and defence.
    - Europe will inevitably have to "chase the hype" in some sectors, mostly B2C ones, in which US and China dominate the scene with very well-established tech giants.

40. Europe features a remarkable leadership on AI research, but is losing ground in global university rankings, and the prospect of Brexit will deprive it of the most vibrant AI research and innovation environment among member states. EU institutions and member states must capitalise on existing knowledge and initiatives to create a new, flourishing environment for AI research in Europe. **If Europe clearly took the leadership on "AI for good", researchers with a strong motivation to develop AI-enabled solutions that address societal challenges may look with more favour at Europe**, especially if they could find a suitable research environment, a well-shaped policy context, and enticing procurement and innovation markets.

41. **The issue of digital skills needs to be addressed, starting with fundamental digital literacy skills, both in education and for research**. At a basic level, digital literacy needs to be taught in schools to enable full participation in a digital society and to achieve the objectives of accessibility for all, agency and human autonomy outlined in the Ethics Guidelines. Interdisciplinary curricula should be developed for all levels of education, bridging computer science with natural sciences and social sciences to equip people with the skills required in an AI-enabled workforce. Basic and applied research should continue to be heavily supported with public funds, coupled with efforts to strengthen Europe's already well-developed AI community, and its relationship with civil society. **Funding should focus on all aspects of the ecosystem, and explore more sustainable, human-centric, privacy-compatible ways of developing AI**. This applies in particular to HPC funding, which should continue and advance on quantum computing, Edge computing and Fog computing; and also to blockchain and smart contracts, as well as the IoT.

42. **Researchers should be given a clearer career path and a smarter set of evaluation criteria**. The creation of spin-offs from university labs, as well as other entrepreneurial initiatives by researchers, should be encouraged as a sign of success. The evaluation of researchers and research teams should not be based only or predominantly on indicators such as the number of patents filed. And universities should be given significantly greater funds in domains such as AI, so that they can compete with the private sector by offering an attractive mix of (greater) intellectual freedom and decent (even if slightly lower) salaries. As an additional element, EU institutions should develop, together with member states, attractive visa programmes for non-EU talent.

43. **A "Mission IT"**: the CEPS Task Force calls for the creation of a catalyst, i.e. an institution or cluster of institutions that leads the European development in this field. A number of activities in that direction have already been started, including CLAIRE, ELLIS, the "HumanE AI" proposed Flagship project and the AI4EU. AI could be featured as an essential component of a future mission to be launched under Horizon Europe. The proposed mission would then have to be broader than AI, embracing the whole technology stack, its ethical, societal and environmental impacts, and the possible consequences for education and skills. In this report, we have provisionally called it "Mission IT".

44. **The CEPS Task Force welcomes the Coordinated Plan and its high level of ambition. The only big element that really seems to be missing is the link between AI and the SDGs**. This is regrettable, since such a link could become the most distinguishing feature of the EU approach to AI and would

establish Europe as a prominent player on AI policy at the global level, including vis-à-vis those countries with which the EU has consolidated trade agreements, or is about to engage in deeper trade negotiations. Taking a clear stance on AI "for good" would also position Europe as a privileged interlocutor of all those private standardisation bodies such as IEEE or ISO, which are extensively working on shaping the development and evolution of AI at the global level; and as an advocate of treating AI for what it is: a means to a more prosperous future, and not a goal in itself.

REFERENCES

Accenture (2018), "An Inclusive Future of Work: A Call to Action" (https://www.accenture.com/t20181114T030204Z__w__/us-en/_acnmedia/PDF-90/Accenture-Inclusive-Future-Of-Work-Full-Report.pdf#zoom=50).

Acemoglu, D. and P. Restrepo (2018a), "Automation and new tasks: The implications of the task content of technology for labour demand", *mimeo*, NBER August.

Acemoglu, D. and P. Restrepo P. (2018b), "Artificial Intelligence, Automation and Work", SSRN Electronic Journal (https://doi.org/10.2139/ssrn.3098384).

Adolphs R. (2015), "The unsolved problems of neuroscience", *Trends in Cognitive Sciences*, 19(4), 173-5.

Aghion, P., Jones B.J., and Jones C. (2017), Artificial Intelligence and Economic Growth. National Bureau of Economic Research, paper no. w23928.

AI Now (2017), "AI Now 2017 Report" (https://ainowinstitute.org/AI_Now_2017_Report.pdf).

Aldewereld, H.M., V. Dignum and Y. Tan (2015), "Design for values in software development", in J. van den Hoven, P.E. Vermaas and I. van de Poel (eds), *Handbook of Ethics, Values, and Technological Design: Sources, theory, values and application domains*, Berlin: Springer.

Armstrong, Mark (2005), "Recent Developments in the Economics of Price Discrimination. Advances in Economics and Econometrics: Theory and Application", Ninth World Congress. 2. 10.1017/CBO9781139052276.006.

Arnold, Thomas, Daniel Kasenberg and Matthias Scheutz (2017), "Value Alignment or Misalignment - What Will Keep Systems Accountable?", paper prepared for AAAI Workshop.

Arthur, W. Brian (1994), *Increasing Returns and Path Dependence in the Economy*, Ann Arbor, MI: University of Michigan Press.

Asilomar AI Principles (2017), Principles developed in conjunction with the 2017 Asilomar Conference [Benevolent AI 2017] (https://futureoflife.org/ai-principles).

Athey, S. (2017), "The impact of machine learning on economics" in Ajay K. Agrawal, Joshua Gans, and Avi Goldfarb (eds), *The Economics of Artificial*

Intelligence: An Agenda Economics of Artificial Intelligence, Chicago, IL: University of Chicago Press.

Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan (2018), "The Moral Machine Experiment", *Nature*, Vol. 563, 59-64.

Azoulay, P., E. Fuchs, A. Goldstein and M. Kearney (2018), "Funding Breakthrough Research: Promises and Challenges of the "ARPA Model", *Innovation Policy and the Economy*, Vol. 19.

Barnet, B.A. (2009), "Idiomedia: The rise of personalized, aggregated content", *Continuum* 23(1): 93–99.

Barocas S and Selbst AD (2015) Big data's disparate impact. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network (http://papers.ssrn.com/ abstract=2477899).

Barocas, S., Moritz Hardt and Arvind Naranayan (2018), *Fairness in Machine Learning* (http://fairmlbook.org).

Bebchuk, Lucian (2001), "Property Rights and Liability Rules: The Ex Ante View of the Cathedral", Harvard Law School John M. Olin Center for Law, Economics and Business Discussion Paper Series, Paper 347.

Berberich, N. and K. Diepold (2018), "Virtous Machine – Old Ethics for New Technology?" (https://arxiv.org/pdf/1806.10322.pdf).

Birrer, F.A.J. (2005), "Data mining to combat terrorism and the roots of privacy concerns", *Ethics and Information Technology* 7(4): 211–220.

Botsman, R. (2017), "Big data meets Big Brother as China moves to rate its citizens", *Wired,* 31 October (https://www.wired.co.uk/article/chinese-government-social- credit-score-privacy-invasion).

Brey, P. and J.H. Soraker (2009), *Philosophy of Computing and Information Technology,* Amsterdam: Elsevier.

Brown, Ian and Christopher Marsden (2013), *Regulating Code: Good Governance and Better Regulation in the Information Age,* Cambridge, MA: MIT Press.

Brundage, M., J. Clark, G.C. Allen, C. Flynn, S. Farquhar, R. Crootof and J. Bryson (2018), "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation", Information Society Project, Future of Humanity Institute (https://www.repository.cam.ac.uk/bitstream/handle/1810/275332/180207228.pdf?sequ ence=1).

Brynjolfsson, E., D. Rock and C. Syverson (2018), "The productivity J-curve: How intangibles complement general purpose technologies", NBER Working Paper No. 25148, NBER, Cambridge, MA, October.

Brynjolfsson, E. and A. McAfee (2014), *The Second Machine Age*, New York and London: W.W. Norton & Co.

Brynjolfsson, E., T. Mitchell and D. Rock (2018), "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?", AEA Papers and Proceedings, Vol. 108 (https://www.aeaweb.org/articles?id=10.1257/pandp.20181019).

Brynjolfsson, E., D. Rock and C. Syverson et al. (2017), "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics".

Cafaggi, F. and A. Renda (2014), "Measuring the Effectiveness of Transnational Private Regulation", October (https://ssrn.com/abstract=2508684).

Cath, C., S. Wachter, B. Mittelstadt et al. (2018), "Artificial Intelligence and the 'Good Society': The US, EU, and UK approach", *Science and Engineering Ethics*, 24 (2): 505-528.

Channon, M., L. McCormick and K. Noussia (2019), *Law and Autonomous Vehicles* Routledge, January.

Charisi, V., C. Liem and E. Gómez (2018), "Novelty-based cognitive processes in unstructured music-making play settings in early childhood", Proceedings of 8th Joint IEEE International Conference of Development and Learning and Epigenetic Robotics ICDL-EpiRob2018, pp. 218-223.

Chen, J., Z. Ju, C. Hua, B. Ma, C. Chen, L. Qin and R. Li (2013), "Accelerated implementation of adaptive directional lifting-based discrete wavelet transform on GPU", *Signal Processing: Image Communication*, 28 (9) pp. 1202–1211.

Chen, N., L. Christensen, K. Gallagher, R. Mate and G. Rafert (2016), "Global economic impacts associated with artificial intelligence", Analysis Group.

Chiacchio, F., G. Petropoulos and D. Pichler (2018), "The impact of industrial robots on EU employment and wages: A local labour market approach", Bruegel Working Paper, Bruegel, Brussels.

Chowdhury, R. (2018), "Tackling the Challenge of Ethics in AI", blog post (https://www.accenture.com/gb-en/blogs/blogs-cogx-tackling-challenge-ethics-ai).

Claffy, K.C. and D. Clark (2014), "Platform models for sustainable Internet regulation", *Journal of Information Policy*, 4, 463–488.

Clark, D. and K.C. Claffy (2015), "Anchoring policy development around stable points: An approach to regulating the co-evolving ICT ecosystem", *Telecommunications Policy* (http://dx.doi.org/10.1016/j.telpol.2015.07.003i).

Cockburn, I.M., R. Henderson and S. Stern (2017), "The Impact of Artificial Intelligence on Innovation", paper prepared for the NBER Conference on Research Issues in Artificial Intelligence, Toronto, September 2017. Forthcoming in *The Economics of Artificial Intelligence*, Ajay Agrawal, Joshua S. Gans and Avi Goldfarb (eds), Chicago, IL: University of Chicago Press.

Commission Nationale Informatique et Liberte (CNIL) (2017), "How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence", Report on the public debate led by the French data protection authority (CNIL) as part of the ethical discussion assignment set by the digital republic bill (https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf).

Committee of the Regions (2017), "Building A European Data Economy", COM(2017) 9 final Brussels.

Conitzer, V (2016), "Philosophy in the face of Artifical Intelligence" (https://arxiv.org/abs/1605.06048).

Cooter, R. and T. Ulen (2004), *Law and Economics* (4th ed.). Reading, MA: Addison Wesley Longman, Inc.

Corea, F. (2018), "AI Knowledge Map: How To Classify AI Technologies", 22 August (https://www.forbes.com/sites/cognitiveworld/2018/08/22/ai-knowledge-map-how-to-classify-ai-technologies/#450891417773).

Council of Europe (CoE) 2017, "Algorithms and Human Rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications" (https://rm.coe.int/study-hr-dimension-of-automated-data- processing-incl-algorithms/168075b94a).

Council of Europe (CoE) 2017, "Technological convergence, artificial intelligence and human rights" (http://www.assembly.coe.int/nw/xml/XRef/Xref-DocDetails-EN.asp?FileID=24236&lang=EN).

Craglia, M., J. Hradec and X. Troussard (2019), "Big Data, AI, and Public Policy" in V. Sucha, D. Mair and M. Sienkiewitz et al. (eds), *Science and Evidence in the Policy Ecosystem Handbook*, Elsevier (forthcoming).

Danaher, J. (2017) "Algocracy as Hypernudging: A New Way to Understand the Threat of Algocracy" (https://ieet.org/index.php/IEET2/more/Danaher20170117).

Darby, Michael R. and Edi Karni (1973), "Free Competition and the Optimal Amount of Fraud", *Journal of Law and Economics*, 16(1): 67–88.

De Vries (2010), "Identity, profiling algorithms and a world of ambient intelligence", *Ethics and Information Technology*, Vol. 12, No.1, pp 71–85.

Delforge, A. and L. Gerard (2017), "Notre vie privée est-elle réellement mise en danger par les robots?: étude des risques et analyse des solutions apportées par le GDPR", *L'intelligence artificielle et le droit*, Brussels: Larcier.

Deloitte (2018), Study on emerging issues of data ownership, interoperability, re-usability and access to data, and liability, prepared for the European Commission (http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51486).

Diakopoulos, N (2015), "Algorithmic accountability: Journalistic investigation of computational power structures", *Digital Journalism* 3(3): 398–415.

Diamond, G.A., B.H. Pollock and J.W. Work (1987), "Clinician decisions and computers", *Journal of the American College of Cardiology* 9(6): 1385–1396.

Dignum et al. (2018), "Ethics by Design. Necessity or Curse?" (http://www.aies/conference.com/2018/contents/papers/main/AIES_2018_paper_68.pdf).

Drexl, J. (2017a), "On the Future EU Legal Framework for the Digital Economy: A Competition-based Response to the Ownership-Access Debate", in S. Lohsse, R. Schulze and D. Staudenmayer (eds), *Trading Data in the Digital Economy: Legal Concepts and Tools*, Baden-Baden: Nomos.

Drexl, J. (2017b), "Designing Competitive Markets for Industrial Data– Between Propertisation and Access", 8, JIPITEC.

Dulleck, Uwe, Matthias Sutter and Rudolf Kerschbamer (2011), "The Economics of Credence Goods: An Experiment on the Role of Liability, Verifiability, Reputation, and Competition", *American Economic Review* 101, 10.1257/aer.101.2.526.

Dupont, B (2016), "Bots, cops, and corporations: on the limits of enforcement and the promise of polycentric regulation as a way to control large-scale cybercrime", *Crime, Law and Social Change,* 2017, Vol. 67, No. 1.

Dutton, T. (2018), "An Overview of National AI Strategies, Medium", 28 June (https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd).

Elsevier (2018), *Artificial Intelligence: How knowledge is created, transferred, and used* (https://www.elsevier.com/research-intelligence/resource-library/ai-report#form).

Etzioni, A. and O. Etzioni (2016), "Designing AI Systems that Obey Our Laws and Values", *Communications of the ACM*, Vol. 59 No. 9.

Eubanks, V. (2018), *Automating Inequality*, London: St. Martin's Press.

European Commission (EC) (1996), Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. (https://eur- lex.europa.eu/lexUriServ/LexUriServ.do?uri =CELEX:31996L0009:EN:HTML).

European Commission (EC) (2003), Directive 2003/98/EC of the European Parliament and of the Council on the re-use of public sector information of 31/12/2003, revised by Directive 2013/37/EC of 17/7/2013.

(https://ec.europa.eu/digital-single- market/en/european-legislation-reuse-public-sector-information).

European Commission (EC) (2016), Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market, COM(2016)593 final, Brussels

European Commission (EC) (2016) Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, OJ L 157.

European Commission (EC) (2016), Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119.

European Commission (EC) (2017), Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the

European Commission (EC) (2017) Joint Communication to the European Parliament and the Council: Resilience, Deterrence and Defence: Building strong cybersecurity for the EU. Tech. rep., EU (https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52017JC0450 &from=ES).

European Commission (EC) (2018), "Proposal for a Council Recommendation on Key Competences for Lifelong Learning." COM(2018) 24 final. Brussels: European Commission. (https://ec.europa.eu/education/sites/education/files/recommendation-key-competences-lifelong-learning.pdf).

European Commission (EC) (2018), AI Factsheet (https://ec.europa.eu/digital-single- market/en/news/factsheet-artificial-intelligence-europe).

European Commission (EC) (2018), Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information, Official Journal of the European Union, L 134/1.

European Commission (EC) (2018), Commission Staff Working Document Guidance on sharing private sector data in the European data economy Accompanying the document Communication from the Commission to the European Parliament, the Council, the European economic and social Committee and the Committee of the Regions "Towards a common European data space", SWD(2018)125 final, Brussels.

European Commission (EC) (2018), Commission Staff Working Document Evaluation of Directive 96/9/EC on the legal protection of databases, SWD(2018) 146 final, Brussels.

European Commission (EC) (2018), Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial Intelligence for Europe COM(2018) 237 final Brussels.

European Commission (EC) (2018), Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, "Towards a common European data space" COM(2018) 232 final Brussels.

European Commission (EC) (2018), Proposal for a Directive of the European Parliament and of the Council on the re-use of public sector information (recast) COM/2018/234 final, Brussels.

European Data Protection Supervisor (EDPS) (2018), Towards a digital ethics, Report from EDPS Ethics Advisory Group (https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf).

European Economic and Social Committee (EESC) (2016), EESC Opinion on Artificial Intelligence (https://www.eesc.europa.eu/en/our-work/opinions-information- reports/opinions/artificial-intelligence).

European Group on Ethics (EGE) (2018), "Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems. European Group on Ethics in Science and New Technologies." Brussels: European Commission (https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf).

European Group on Ethics in Science and New Technologies (2018), Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems, March (https://ec.europa.eu/info/news/ethics-artificial-intelligence-statement-ege-released-2018-apr- 24_en).

European Parliament (EP) (2016), "European Civil Law Rules in Robotics".

European Parliament (EP) (2017), Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL).

Federal Trade Commission (2016), Big Data A Tool for Inclusion or Exclusion? Understanding the Issues, FTC Report, January.

Feijoo, C. et al. (2019), "The Industrial Innovation Ecosystem of Artificial Intelligence in China: Current status and prospects", European Commission, Joint Research Centre, Seville, Spain (forthcoming).

Ferguson, A.G. (2017), *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, New York, NY: NYU Press.

Fiorillo, G., Paolo Bocchini and Javier Buceta (2018), "A Predictive Spatial Distribution Framework for Filovirus-Infected Bats", *Scientific Reports*, 2018; 8 (1) DOI: 10.1038/s41598-018-26074-4

Floridi, L. (2016), "Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions" (http://rsta.royalsocietypublishing.org/content/374/2083/20160112).

Floridi, L. (2008), "The method of levels of abstraction", *Minds and Machines,* 18(3): 303–329.

Floridi, L. (2011), "The informational nature of personal identity", *Minds and Machines* 21(4): 549–566.

Floridi L (2012), Big data and their epistemological challenge. Philosophy and Technology 25(4): 435–437.

Floridi, L. (2013), *The Ethics of Information,* Oxford: Oxford University Press.

Floridi, L. (2018), "Soft ethics and the governance of the digital"*, Philosophy and Technology*, 31 (https://doi.org/10.1007/s13347-018-0303-9).

Floridi, L. (ed.) (2015), *The Onlife Manifesto: Being Human in a Hyperconnected Era,* London: Springer Open.

Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke and E. Vayena (2018), "AI4People White Paper: Twenty Recommendations for an Ethical Framework for a Good AI Society", forthcoming in *Minds and Machines*, December.

France Intelligence Artificielle (FIA) (2017), Rapport de synthèse. (https://www.economie.gouv.fr/files/files/PDF/2017/Rapport_synthese_France_IA_.pdf).

France Intelligence Artificielle (FIA) (2017). Rapport de synthèse. Groupe de travail.

Freedman, R., J. Schaich Borg, W. Sinnott-Armstrong, J. Dickerson and V. Conitzer (2018), "Adapting a Kidney Exchange Algorithm to Align with Human Values" in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA.

Frey, C.B. and M.A. Osborne (2013), "The Future of Employment", Working Pape. Oxford Martin Programme on Technology and Employment. (https://www.oxfordmartin.ox.ac.uk/downloads/academic/future-of-employment.pdf).

Frey, C.B. and M.A. Osborne (2017), "The Future of Employment: How Susceptible Are Jobs to Computerisation?", *Technological Forecasting and Social Change*, 114: 254–80 (https://doi.org/10.1016/j.techfore.2016.08.019).

Fuchs, E.R.H. (2010), “Rethinking the role of the state in technology development: DARPA and the case for embedded network governance”, *Research Policy* 39, 1133–1147.

Furman, J. and R. Seamans (2018), “AI and the Economy”, NBER Working Paper No. 24689, NBER, Cambridge, MA.

Future of Privacy Forum (2017), “Unfairness by Algorithm: Distilling the Harm of Automated Decision-Making” (https://fpf.org/2018/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/).

Future Society (2017,) “Making the AI revolution work for everyone” (https://www.tuftsgloballeadership.org/sites/default/files/images/resources/Miailhe%20Reading.pdf).

Geist, Edward and Andrew J. Lohn, (2018), “How Might Artificial Intelligence Affect the Risk of Nuclear War?”, Santa Monica, CA: RAND Corporation (https://www.rand.org/pubs/perspectives/PE296.html).

Giuffrida, I., F. Lederer and N. Vermerys (2018), “A Legal Perspective on the Trials and Tribulations of AI: How Artificial Intelligence, the Internet of Things, Smart Contracts, and Other Technologies Will Affect the Law”, 68 *Case W. Res. L. Rev.* 747 (https://scholarlycommons.law.case.edu/caselrev/vol68/iss3/14).

Goldfarb, A., D. Trefler (forthcoming 2019), “AI and International Trade”, in A. Agrawal, J.S. Gans and A. Goldfarb (eds), *The Economics of Artificial Intelligence,* Chicago, IL: University of Chicago Press.

Gomez-Uribe, C.A. and N. Hunt (2016), “The Netflix recommender system: Algorithms, business value, and innovation”, *ACM Transactions on Management Information Systems*, Vol. 6, No. 4, January.

Graetz, G. and G. Michaels (2017), “Is Modern Technology Responsible for Jobless Recoveries?”, *American Economic Review*, 107(5), 168–73.

Graetz, G. and G. Michaels (2018), “Robots at Work”, *Review of Economics and Statistics*, Vol. 100, 753-768.

Gregory, T., A. Salomons and U. Zierahn (2018), “Racing with or against the machine, evidence from Europe”, CESifo Working Paper No. 7247, September.

Grindro,d P. (2014), *Mathematical Underpinnings of Analytics: Theory and Applications,* Oxford: OUP.

Guillén, M. and S. Reddy (2018), “We know ethics should inform AI. But which ethics?”, World Economic Forum note (https://www.weforum.org/agenda/2018/07/we-know-ethics- should-inform-ai-but-which-ethics-robotics/).

Hall, W. and J. Pesenti (2017). Growing the Artificial Intelligence Industry in the UK (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/ Growing_the_artificial_intelligence_industry_in_the_UK.pdf).

Hardt, M., E. Price and N. Srebro (2016), "Equality of opportunity in supervised learning", *Advances in neural information processing systems* (pp. 3315-3323) (https://arxiv.org/abs/1610.02413).

Hildebrandt, M. (2008), "Defining profiling: A new type of knowledge?", in M. Hildebrandt and S. Gutwirth (eds), *Profiling the European Citizen*, Springer (http://link.springer.- com/chapter/10.1007/978-1-4020-6914-7_2).

Hind, M., S. Mehta, A. Mojsilovic, R. Nair, K. Natesan Ramamurthy, A. Olteanu, and K.R. Varshney (2018), "Increasing Trust in AI Services through Supplier's Declarations of Conformity" (https://arxiv.org/pdf/1808.07261.pdf).

House of Lords Artificial Intelligence Committee (2018) "AI in the UK: ready, willing and able?", 16 April (https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm.

Huang, L., A. Joseph, B. Nelson et al. (2011), "Adversarial Machine Learning", in Proceedings of 4th ACM Workshop on Artificial Intelligence and Security, October 2011 (https://people.eecs.berkeley.edu/~tygar/papers/SML2/Adversarial_AISEC.pdf).

Hubbard, F.P. (2011), "Do Androids Dream? Personhood and Intelligent Artifacts", 83 *Temp. L. Rev*. 405.

Hutchinson, Ben and Margaret Mitchell (2019), "50 Years of Test (Un)fairness: Lessons for Machine Learning", in Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). ACM, New York, NY, DOI: https://doi.org/10.1145/3287560.3287600.

IEEE (2018), "Ethically Aligned Design" (http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html).

IEEE (2017), "Ethically Aligned Design, v2", Initiative on Ethics of Autonomous and Intelligent Systems (https://ethicsinaction.ieee.org).

IEEE (2017), "Special Issue: Hacking without Humans. IEEE Security & Privacy" (https://publications.computer .org/security-and-privacy/tag/hacking-without-humans/).

International Labour Organisation (ILO) (2019), "Work for A Brighter Future", Report of the Global Commission on the Future of Work, Geneva.

Jasanoff, S. (2013), "Technologies of humility: Citizen participation in governing science", *Minerva*, 41(3): 223-244.

Johnson, J.A. (2013), "Ethics of data mining and predictive analytics in higher education", SSRN Scholarly Paper, Rochester, NY: Social Science Research Network (http://papers.ssrn.com/abstract=2156058).

JRC (European Commission Joint Research Centre) (2018), "Artificial intelligence: A European Perspective", Flagship Report on AI.

Kerber, W. (2016), "A new intellectual property right for data: an economic analysis". GRUR 989.

Kerber, W. (2016), "Governance of Data: Exclusive Property vs. Access", *International Review of Intellectual Property and Competition Law*, 47: 7, 759–762 (https://link.springer .com/article/10.1007/s40319-016-0517-2).

Kerber, W. and J.S. Frank (2017), "Data Governance regimes in the Digital Economy: The Example of Connected Cars" (https://www.tilburguniversity.edu/upload/5c3e205b-fc0e-42ef-a1e8-8d72c0e16025_Kerber_Frank_01102017_Data%20Governance%20and%20Connected% 20Cars.pdf).

Kim, H, J. Giacomin and R. Macredie (2014), "A qualitative study of stakeholders' perspectives on the social network service environment", *International Journal of Human–Computer Interaction* 30(12): 965–976.

Kingston, John (2017), "Using Artificial Intelligence to Support Compliance with the General Data Protection Regulation", *Artificial Intelligence and Law* 1-15. 10.1007/s10506-017-9206-9.

Korinek, A. and J. Stiglitz (2017), "Artificial Intelligence and Its Implications for Income Distribution and Unemployment" (https://doi.org/10.3386/w24174).

Lee, June-Goo, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo and Namkug Kim (2017), "Deep Learning in Medical Imaging: General Overview", *Korean Journal of Radiology*, 18 (570. 10.3348/kjr.2017.18.4.570).

Lessig, L. (1999), *Code and other Laws of Cyberspace*, New NY: Basic Books.

Listokin, Y. (2008), "Learning through Policy Variation", 118 *YALE L.J.*

Lomborg S. and A. Bechmann (2014), "Using APIs for data collection on social media", *Information Society* 30(4): 256–265.

Mahajan, R.L., J. Reed, N. Ramakrishnan et al. (2012), "Cultivating emerging and black swan technologies", ASME 2012 International Mechanical Engineering Congress and Exposition, Houston, TX, pp. 549–557.

Manca A.R., P. Benczur and E. Giovannini (2017), "Building a Scientific Narrative Towards a More Resilient EU Society", JRC Science for Policy Report (http://publications.jrc.ec.europa.eu/repository/bitstream/JRC106265/jrc106265_100417 _resilience_scienceforpolicyreport.pdf).

Marsden, Christopher (2011), "Internet Co-Regulation and Constitutionalism: Towards a More Nuanced View?, 29 August (SSRN, https://ssrn.com/abstract=1973328).

Mazoue, J.G. (1990), "Diagnosis without doctors", *Journal of Medicine and Philosophy* 15(6): 559–579.

McKinsey Global Institute (2017), "A future that works: Automation, employment and productivity".

Mendoza, I. and L.A. Bygrave (2017), "The Right Not to Be Subject to Automated Decisions Based on Profiling", in T.E. Synodinou et al. (eds), *EU Internet Law: Regulation and Enforcement*, Springer.

Micheler, E. and A. Whaley (2018), "Regulatory Technology: Replacing Law with Computer Code", LSE Law, Society and Economy Working Papers 14/2018 London School of Economics and Political Science Law Department.

Miller, C., J. Ohrvik-Stott and R. Coldicutt (2018), *Regulating for Responsible Technology: Capacity, Evidence and Redress: A new system for a fairer future*, London: Doteveryone (https:// doteveryone.org.uk/project/regulating-for-responsible-technology).

Mission Villani sur l'intelligence artificielle (2018), "For a meaningful artificial intelligence. Towards a French and European Strategy" (https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf).

Mittelstadt, B.D. and L. Floridi (2016), "The ethics of big data: Current and foreseeable issues in biomedical contexts", *Science and Engineering Ethics* 22(2): 303–341.

Mittelstadt, B.D., P. Allo, M. Taddeo, S. Wachter and L. Floridi (2016), "The Ethics of Algorithms: Mapping the Debate", *Big Data & Society,* July–December, 1–21.

Montreal Declaration (2017), For a Responsible Development of Artificial Intelligence (https://www.montrealdeclaration-responsibleai.com/the-declaration).

Nabi, J. (2018), "How Bioethics Can Shape Artificial Intelligence and Machine Learning", Hastings Center Report 48, no. 5 (2018): 10–13. DOI: 10.1002/hast.895

Narayanan, A. (2018), "Fat* tutorial: 21 fairness definitions and their politics", New York, NY.

National Science and Technology Council (NSTC) (2016a), "Preparing for the future of artificial intelligence", Executive Office of the President (https://info.publicintelligence.net/WhiteHouse-ArtificialIntelligencePreparations.pdf).

Nelson, Phillip (1970), "Information and Consumer Behavior", *Journal of Political Economy*, 78(2): 311–329.

Nevejans, N. (2016), "Civil Law Rules for Robotics", study for the European Parliament JURI Committee, PE 571.379 (http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/ IPOL_STU(2016)571379_EN.pdf).

Newell, S. and M. Marabelli (2015), "Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'", Journal of Strategic Information Systems 24(1): 3–14.

Nilsson, N.J. (2010), *The Quest for Artificial Intelligence: A History of Ideas and Achievements,* Cambridge: Cambridge University Press.

Noble, S.U. (2018), *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, NY: New York University Press.

O'Neil, C. (2016), *Weapons of math destruction: How big data increases inequality and threatens democracy*, New York, NY: Crown Publishers.

OECD (2018), *Job Creation and Local Economic Development: Preparing for the Future of Work*, OECD Publishing, Paris (https://doi.org/10.1787/9789264305342-en).

Office parlementaire d'évaluation des choix scientifiques et technologiques (OPECST) (2017), "Toward a Controlled, Useful and Demystified Artificial Intelligence" (http://www.senat.fr/rap/r16-464-1/r16-464-1-syn-en.pdf).

Osborne, Clarke LLP (2016), "Legal study on Ownership and Access to Data", European Commission (https://publications.europa.eu/en/publication-detail/publication/d0bec895-b603-11e6-9e3c-01aa75ed71a1/language-en).

Palacin, M., M. Oliver, J. Infante, S. Oechsner and A. Bikfalvi (2013), "The Impact of Content Delivery Networks on the Internet Ecosystem", *Journal of Information Policy*, Vol. 3, pp. 304-330.

Partnership on AI (2018), Tenets.

Pasquale, F. (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge, MA: Harvard University Press.

Pau J., J. Bajer and N. Houston (2017), "Artificial intelligence in Asia: Preparedness and Resilience", Asia Business Council (http://www.asiabusinesscouncil.org/Research17_AI.html).

Paxon, V. (1998). "Bro: A System for Detecting Network Intruders in Real-Time. Computer Networks" (https://www.usenix.org/legacy/publications/library/proceedings/sec98/full_papers/paxso n/paxson.pdf).

Penin, J. (ed.) (2018), *Intellectual Property and Digital Trade in the Age of Artificial Intelligence and Big Data* (https://www.ictsd.org/themes/innovation-and-ip/research/intellectual-property-and-digital-trade-in-the-age-of-artificial).

Portmess, L. and S. Tower (2014), "Data barns, ambient intelligence and cloud computing: The tacit epistemology and linguistic representation of Big Data", *Ethics and Information Technology* 17(1): 1–9.

Purdy, M. and P. Dougherty (2017), "Why Artificial Intelligence is the Future of Growth", Accenture/Frontier Economics Report (https://www.accenture.com/t20170927T080049Z__w__/us-en/_acnmedia/PDF-33/Accenture-Why-AI-is-the-Future-of-Growth.PDFla=en).

PWC (2017), "Sizing the Prize. What's the real value of AI for your business and how can you capitalise?", PWC Analysis (https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf).

Ranchordás, S. (2013), "Experimental Legislation: The Whys and the Woes", 1 *Theory & Prac. Legis*. 415.

Rathenau Institute (2017), "Human rights in the robot age", Council of Europe Report (https://www.rathenau.nl/sites/default/files/2018-02/Human%20Rights%20in%20the%20Robot%20Age-Rathenau%20Instituut-2017.pdf).

Raymond, A. (2014), "The dilemma of private justice systems: Big Data sources, the cloud and predictive analytics", *Northwestern Journal of International Law & Business*, forthcoming (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2469291).

Reisman, D., J. Schultz, K. Crawford and M. Whittaker (2018), "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability", AI Now Institute, New York University.

Renda, A. (2015), "Searching for Harm, or Harming Search? A look at the European Commission's antitrust investigation against Google", CEPS Special Report No. 118/September. Published also as Working Paper of the Rethinking Regulation programme at Duke University, Kenan Institute for Ethics.

Renda, A. (2016), "Selecting and Designing European ICT Innovation Policies", Report for the European Commission, Joint Research Centre and Institute for Prospective Technological Studies, September.

Renda, A. (2017a), "How can Sustainable Development Goals be 'mainstreamed' in the EU's Better Regulation Agenda?", CEPS Policy Insights No 2017/12, CEPS, Brussels, March.

Renda, A. (2017b), “Reforming E-Communications Services”, Critical Assessment Report for the European Parliament IMCO Committee (https://www.ceps.eu/system/files/IPOL_IDA%282017%29595348_EN.pdf).

Renda, A. (2017c), “Will the DSM Strategy Spur Innovation?”, *Intereconomics*, Vol. 52, No. 4, pp 197-201.

Renda, A. (2018a), “Ethics, Algorithms and Self-Driving Cars: A CSI of the ‘Trolley Problem’”, CEPS Policy Insights No. 2018/02, January (https://www.ceps.eu/system/files/PI%202018-02_Renda_TrolleyProblem.pdf).

Renda, A. (2018b), “Cost-Benefit Analysis and EU Policy”, in S. Garben and I. Govaere (eds), The EU Better Regulation Agenda: A Critical Assessment, Portland, OR: Hart Publishing.

Renda, A. (2018c), “The legal framework to address “fake news”: Possible policy actions at the EU level”, Report for the European Parliament, IMCO Committee, Policy Department for Economic, Scientific and Quality of Life Policies, Directorate-General for Internal Policies, PE 619.013- June.

Renda, A. (2019), “Up, down, and sideways: the endless quest for EU’s optimal multi-level governance”, forthcoming in E. Brousseau, J.M. Glachant and J. Sgard (eds), *Oxford Handbook of International Economic Governance and Market Regulation*, Oxford: Oxford University Press.

Renda, A. and C. Yoo (2015), “Telecommunications and Internet Services: The digital side of the TTIP”, Paper No. 8 in the CEPS-CTR project TTIP in the Balance and CEPS Special Report No. 112, July 2015, now published as Chapter 12 in D. Hamilton and J. Pelkmans (2015), *Rule-Makers or Rule-Takers? Exploring the Transatlantic Trade and Investment Partnership*, London: Rowman & Littlefield.

Posner, R.A. and W.M. Landes (1980), "The Positive Economic Theory of Tort Law," 15 *Georgia Law Review* 851.

Russell, S. and P. Norvig (2009), *Artificial Intelligence: A Modern Approach* (3rd Edition), Pearson.

Russell, Stuart, Daniel Dewey and Erik Max Tegmark (2015), “Research Priorities for Robust and Beneficial Artificial Intelligence”, *AI Magazine*, 36, pp. 105-114.

Shapiro, C. and H. Varian (1998), *Information Rules. A Strategic Guide to the Network Economy*, Cambridge, MA: Harvard Business School Press.

Silver, D., Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine

Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis, "Mastering the game of Go with deep neural networks and tree search", *Nature* 529 (2016): 484–489.

Simon, H.A. (1971), "Designing Organizations for an Information-Rich World", in Martin Greenberger (ed.), *Computers, Communication, and the Public Interest,* Baltimore, MD: The Johns Hopkins Press, pp. 40–41.

Solum, L.B. (1992), "Legal Personhood for Artificial Intelligences", 70 *N.C. L. Rev*. 1231.

Stam, K. et al. (2015), "Employment status and subjective well-being: The role of the social norm to work", *Work, employment and society*, 1–25.

Stevenson, B. (2017), "AI, Income, Employment, and Meaning" in A. Agrawal, J.S. Gans and A. Goldfarb (eds), *The economics of artificial intelligence,* Chicago, IL: University of Chicago Press (http://www.nber.org/chapters/c14026).

Sunstein, C. and R.H. Thaler (2008), *Nudge. Improving Decisions about Health, Wealth, and Happiness*, Penguin Books.

Sweeney (2013), "Discrimination in Online Ad Delivery", Communications of the Association of Computing Machinery (CACM).

Syverson, Chad (2017), "Challenges to Mismeasurement Explanations for the US Productivity Slowdown", *Journal of Economic Perspectives,* 31 (2): 165-186.

Tegmark, M. (2017), *Life 3.0, Being Human in the Age of Artificial Intelligence*, New York, NY: Random House.

Tickle, A.B. et al. (1998), "The Truth Will Come to Light: Directions and Challenges in Extracting the Knowledge Embedded within Trained Artificial Neural Networks", *IEEE Trans. Neural Netw.*, Vol. 9, No. 6, pp. 1057–1068.

Tutt, Andrew (2017), "An FDA for Algorithms" (March 15, 2016), 69 *Admin. L. Rev*. 83.

UK Government Office for Science (2016), "Artificial intelligence: opportunities and implications for the future of decision making", (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/ gs-16-19-artificial-intelligence-ai-report.pdf).

UK Government (2018), "Industrial Strategy: Artificial Intelligence Sector Deal", Policy Paper, 26 April (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/702810/180425_BEIS_AI_Sector_Deal__4_.pdf).

UK House of Commons Science and Technology Committee (2016), "Robotics and Artificial Intelligence", (https://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf).

UK House of Lords (2017), "AI in the UK: ready, willing and able?", Select Committee on Artificial Intelligence Report of Session 2017-19, HL Paper 100.

Van Dijck, J. (2014), "Datafication, dataism and dataveillance: Big data between scientific paradigm and ideologue", *Surveillance & Society*, 12 (2): 197-208.

Veale, M. and L. Edwards (2018), "Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling", *Computer Law & Security Review*, Vol. 34, No. 2, 2018; DOI: 10.1016/j.clsr.2017.12.002.

Verma, S. and J. Rubin (2018), "Fairness definitions explained", 1-7. 10.1145/3194770.3194776.

Wachter, S. et al. (2017), "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law*, Vol. 7, No. 2, pp. 76–99; DOI: 10.1093/idpl/ipx005.

Wachter, S., B. Mittelstadt and C. Russell (2017), "Counterfactual explanations without opening the black box: Automated decisions and the GDPR" (https://arxiv.org/abs/1711.00399).

Whitt, R.S. (2007), "Adaptive policymaking: Evolving and applying emergent solutions for U.S. communications policy", *Federal Communications Law Journal*, 61(3), 483-589.

World Economic Forum (2018), "Harnessing Artificial Intelligence for the Earth", Report in Collaboration with PwC and Stanford Woods Institute for the Environment, January (http://www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the_Earth_report_2018.pdf).

Xi Jinping (2014), The Governance of China.

Yeung, K. (2017a), "'Hypernudge': Big data as a mode of regulation by design", *Information, Communication and Society*, Vol. 20, No. 1, The Social Power of Algorithms.

Yeung, K. (2017b), "Algorithmic Regulation: A Critical interrogation", *Regulation and Governance,* doi: 10.1111/rego.12158.

Yeung, K. (2008), "Towards an understanding of regulation by design", in R. Brownsword and K. Yeung (eds), *Regulating technologies,* Portland, OR: Hart Publishing.

Yeung, K. and M. Dixon-Woods (2010), "Design-based regulation and patient safety: A regulatory studies perspective", *Social science & medicine*. 71. 502-9. 10.1016/j.socscimed.2010.04.017.

Yi Cheng, Kate Yen, Yeqi Chen, Sijin Chen and Alexis Hiniker (2018), "Why doesn't it work?", in Proceedings of the 17th Interaction Design and Children Conference, 2018 DOI: 10.1145/3202185.3202749.

ANNEX

CEPS TASK FORCE MEMBERS AND GUEST SPEAKERS

Chair: Andrea Renda, CEPS Senior Research Fellow and Head of Global Governance, Regulation, Innovation and the Digital Economy

Task Force Members

Dilek Ayayadin, Executive Administrator, Luxembourg House of Financial Technology

Ginevra Bruzzone, Deputy Director General, Assonime and Senior Research Fellow, School of European Political Economy, Luiss

Alessandra Casale, Head of Brussels Office, Assonime

Eve Chen, Regulatory Strategy and Government Relations, London Stock Exchange Group

Elisabeth Crossick, Head of Government Relations, RELX Group

Frederick De Backer, Manager Regulatory affairs, Telefonica

Tim Dutz, Manager Strategic Projects, Schufa

Olof Erixon, Senior Legal Counsel, Confederation of Swedish Enterprises

Eleanor Flanagan, Senior Manager, Regulatory Affairs, Spotify

Brit Hecht, Head of Digital Public Affairs, BBVA

Nicholas Hodac, Government and Regulatory Affairs Executive, IBM Europe

Grania Holzwarth, Legal Counsel, Deutsche Telekom

Elif Kiesow Cortez, Senior Researcher and Lecturer, The Hague University of Applied Sciences

Tomislav Korman, Public Relations Officer, European Parliament

Aleksandra Kozik, Head of Stakeholder Engagement, EU Affairs, Facebook

Cornelia Kutterer, Senior Director, EU Government Affairs, Privacy and Digital Policies, Microsoft

Danai Lazari, Junior Consultant, Afore Consulting

Eva Lichtenberger, former MEP, European Parliament

Guido Lobrano, Vice President Europe, Information Technology Industry Council

Jesus Lozano Belio, Senior Economist in Digital Regulation and Trends, BBVA

Christph Luykx, Chief Privacy Strategist, CA Technologies

Manuela Mackert, Chief Compliance Officer, Deutsche Telekom

Riccardo Masucci, Global Director of Privacy Policy, Intel

Cosmin Mircea, Public Policy and Government Affairs Analyst, RELX Group

Monica Monaco, Founder and Managing Director, TrustEUAffairs

Susana Nascimento, Policy Analyst, Joint Research Centre, European Commission

Alex Panican, Head of Partnerships and Ecosystems, Luxembourg House of Financial Technology

Agnieszka Papaj, EU Policy Manager for Poland and Central Europe, Deloitte

Emmanouil Patavos, Senior Director, Strategic Communications, FTI Consulting

Carlos Rodriguez Cocina, Director European Regulatory Affairs and Head of Brussels Office, Telefonica

Mario Romao, Global Director, Health and Data Policy, Intel

Ondrej Socuvka, Senior Public Policy and Government Affairs Manager, Google

Tim Stok, Government Affairs Manager, RELX Group

Andreas Tegge, Head of Global Public Policy, SAP

Stefan Van Duin, Partner Analytics & Information Management, Deloitte

Lucia Vesnic-Alujevic, Policy Analyst, Joint Research Centre, European Commission

Elodie Vignon, Senior Consultant Public Affairs, Zurich Insurance

Daniel Voelsen, Research Associate, SWP Berlin

Barbara Wynne, Director EU government Relations, Accenture

Guest Speakers

Michal Boni, MEP, European Parliament

Daniel Braun, Deputy Head of Cabinet of Commissioner for Justice, European Commission

Aljoscha Burchardt, Senior Researcher and Lab Manager, German Research Center for Artificial Intelligence (DFKI)

Maria Chiara Carrozza, Professor, Scuola Superiore Sant’Anna, Pisa

Rumman Chowdhury, Responsible AI Lead, Accenture

Vincent Conitzer, Professor of New Technologies, Duke University

Nathalie Devillier, Grenoble School of Management, Member of the European Commission High Level Group on Liability and New Technologies

Marcel Dickow, Head of Research Division, SWP Berlin

Jim Dratwa, European Group on Ethics in Science, European Commission

Iordana Eleftheriadou, Head of Advanced Tecnologies and Digital Transformation Team, European Commission

Luciano Floridi, Professor of Philosophy and Ethics of Information, Oxford Internet Institute

Dean Garfield, President and CEO, Information Technology Industry Council

Paolo Grassia, Head of Regulation and Advocacy, European Telecommunications Network Operators

Daniel Voelsen, Research Associate, SWP Berlin

Rina Joosten, Co-Founder and CCO, Seedlink

Bjoern Juretzki, Assistant to the Director-General, DG CONNECT, European Commission

Lenard Koschwitz, Senior Director Global Public Policy, Allied for Startups

Mark Nitzberg, Executive Director, UC Berkeley Center for Human-Compatible AI

Michal Pěchouček, Professor of Computer Science, Czech Technology University

Fabrizio Porrino, SVP Global Public Affairs, FacilityLive

Aleksandra Przegalińska, Assistant Professor, Kozminski University/ Management in Netrworked and Digital SocietiesM and Research Fellow, MIT Center for Collective Intelligence

Francesca Rossi, AI Ethics Global Leader, IBM and Professor, University of Padova

Koen Simoens, VP operations, Sentiance

Philipp Slusallek, Scientific Director, German Research Center

for Artificial Intelligence (DFKI) and Professor, Saarland University

Toby Walsh, Professor of Artificial Intelligence, UNSW Sydney and Technical University of Berlin

John Zysman, Professor of Political Science, UC Berkeley Center for Human-Compatible AI

Despite still being in its infancy, artificial intelligence (AI) has already shown enormous potential to advance humanity towards new frontiers of prosperity and growth. Due to its powerful force, however, AI can also pose significant risks, which should be handled with the utmost care, but not with fear. The world's top political leaders have understood AI's disruptive potential and are rushing to secure a competitive advantage in this crucial emerging domain, even at the price of reviving old-fashioned industrial policy. At the same time, academia and civil society are calling for widely shared ethical principles to avoid negative repercussions. In this fast-changing context, Europe is struggling to keep pace with superpowers like the United States and China.

This report summarises the work of the CEPS Task Force on Artificial Intelligence, which met throughout 2018. Arguing that the EU and its member states are uniquely positioned to lead the global community towards responsible, sustainable AI development, its members call upon European leaders to focus on leveraging AI's potential to foster sustainable development, in line with the future 2030 Agenda. The report puts forward 44 recommendations on how to design and promote lawful, responsible and sustainable AI and how to approach future policy and investment decisions with the aim of positioning Europe in the driver's seat to address the most disruptive technology transition of our times.

9 789461 387165

CEPS