Artificial intelligence: How does it work, why does it matter, and what can we do about it?

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Author: Philip Boucher
Scientific Foresight Unit (STOA)
PE 641.547 – June 2020

EN

Artificial intelligence: How does it work, why does it matter, and what can we do about it?

Artificial intelligence (AI) is probably the defining technology of the last decade, and perhaps also the next. The aim of this study is to support meaningful reflection and productive debate about AI by providing accessible information about the full range of current and speculative techniques and their associated impacts, and setting out a wide range of regulatory, technological and societal measures that could be mobilised in response.

AUTHOR

Philip Boucher, Scientific Foresight Unit (STOA),

This study has been drawn up by the Scientific Foresight Unit (STOA), within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

To contact the publisher, please e-mail stoa@ep.europa.eu

LINGUISTIC VERSION

Original: EN

Manuscript completed in June 2020.

PE 641.547
ISBN: 978-92-846-6770-3
doi: 10.2861/44572
QA-01-20-338-EN-N

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)

Executive summary

Artificial intelligence (AI) is probably the defining technology of the last decade, and perhaps also the next. The aim of this study is to support meaningful reflection and productive debate about AI by providing accessible information about the full range of current and speculative techniques and their associated impacts, and setting out a wide range of regulatory, technological and societal measures that could be mobilised in response.

What is artificial intelligence?

The study adopts the European Commission's 2018 definition of AI, which is both accessible and typical of contemporary definitions.

> AI refers to systems that display intelligent behaviour by analysing their environment and taking action – with some degree of autonomy – to achieve specific goals.

Since AI refers to so many techniques and contexts, greater precision is required in order to hold meaningful and constructive debates about it. For example, arguments about simple 'expert systems' used in advisory roles need to be distinguished those from those concerning complex data-driven algorithms that automatically implement decisions about individuals. Similarly, it is important to distinguish arguments about speculative future developments that may never occur from those about current AI that already affects society today.

How does artificial intelligence work?

Chapter 2 sets out accessible introductions to some of the key techniques that come under the AI banner. They are grouped into three sections, which gives a sense of the chronology of the development of different approaches.

The first wave of early AI techniques is known as 'symbolic AI' or expert systems. Here, human experts create precise rule-based procedures – known as 'algorithms' – that a computer can follow, step by step, to decide how to respond intelligently to a given situation. Fuzzy logic is a variant of the approach that allows for different levels of confidence about a situation, which is useful for capturing intuitive knowledge, so that the algorithm can make good decisions in the face of wide-ranging and uncertain variables that interact with each other. Symbolic AI is at its best in constrained environments which do not change much over time, where the rules are strict and the variables are unambiguous and quantifiable. While these methods can appear dated, they remain very relevant and are still successfully applied in several domains, earning the endearing nickname 'good old-fashioned AI'.

The second wave of AI comprises more recent 'data-driven' approaches which have developed rapidly over the last two decades and are largely responsible for the current AI resurgence. These automate the learning process of algorithms, bypassing the human experts of first wave AI. Artificial neural networks (ANNs) are inspired by the functionality of the brain. Inputs are translated into signals which are passed through a network of artificial neurons to generate outputs that are interpreted as responses to the inputs. Adding more neurons and layers allow ANNs to tackle more complex problems. Deep learning simply refers to ANNs with several layers. Machine learning (ML) refers to the transformation of the network so that these outputs are considered useful – or intelligent – responses to the inputs. ML algorithms can automate this learning process by making gradual improvements to individual ANNs, or by applying evolutionary principles to yield gradual improvements in large populations of ANNs.

The third wave of AI refers to speculative possible future waves of AI. While first and second wave techniques are described as 'weak' or 'narrow' AI in the sense that they can behave intelligently in specific tasks, 'strong' or 'general' AI refers to algorithms that can exhibit intelligence in a wide range of contexts and problem spaces. Such artificial general intelligence (AGI) is not possible with current technology and would require paradigm shifting advancement. Some potential approaches have been considered, including advanced evolutionary methods, quantum computing and brain emulation. Other forms of speculative future AI such as self-explanatory and contextual AI can seem modest in their ambitions, but their potential impact – and barriers to implementation – should not be underestimated.

Why does artificial intelligence matter?

Chapter 3 builds upon the understanding of how these technologies work to examine several opportunities and challenges presented by their application in various contexts.

Several challenges are associated with today's AI. Broadly, they can be understood as a balancing act between avoiding underuse whereby we miss out on potential opportunities, and avoiding overuse whereby AI is applied for tasks for which it is not well suited or results in problematic outcomes. The ML process makes some algorithms vulnerable to bias, and their complexity makes their decision-making logic difficult to understand and explain. There are some important challenges in ensuring that the costs and benefits of AI development are distributed evenly, avoiding the concentration of resources in uncompetitive markets and prioritising applications that alleviate rather than exacerbate existing structural inequalities. Other key challenges include the public acceptability of the technology, its alignment with social values, and concerns about some military applications.

There are also several longer-term opportunities and challenges that are contingent upon future developments which might never happen. Some utopian and dystopian scenarios might contribute to hype cycles, but they also present an opportunity to prepare for more moderate trends and reflect upon what we want from the technology. For example, it has been suggested that AI could lead to major job losses or make the concept of employment obsolete, that it could escape human control and take control of its own development, that it could challenge human autonomy or develop artificial emotions or consciousness, presenting interesting – yet speculative – philosophical questions.

What can we do about artificial intelligence?

Chapter 4 sets out several options that could be mobilised in response to the opportunities and challenges that were set out in the previous chapter. The options are organised into three sections, focussing on policy, technology and society. Each section contains seven themes with several options for each, with over 100 measures in total.

Most AI policy debates concern how to shape the regulatory and economic context in which AI is developed and applied in order to respond to specific opportunities and challenges. These could include creating a supportive economic and policy context, promoting more competitive ecosystems, improving the distribution of benefits and risks, building resilience against a range of problematic outcomes, enhancing transparency and accountability, ensuring mechanisms for liability and developing governance capacity.

There are also more abstract policy debates about the broad regulatory approach. This includes questions about whether to have regulation that specifically targets AI, or to regulate it by applying and updating tech-neutral mechanisms that apply to all activities, regardless of whether they use AI.

Similarly, there are institutional debates about whether to set up dedicated institutions for AI, or to make use those that we already have. Another broad question concerns where to regulate, e.g. at Member State level, European Union (EU) level, through other institutions such as the Organisation for Economic Co-operation and Development (OECD) and United Nations (UN), or via self-regulation by actors in the AI sector.

It is also possible to shape the development and application of AI through technological actions. They could include measures related to technology values, the accessibility and quality of data and algorithms, how applications are chosen and implemented, the use and further development of 'tech fixes', and encouraging more constructive reflection and critique.

Finally, societal and ethics measures could be taken, targeting the relationship between AI and society, taking account of social values, structures and processes. These could include measures related to skills, education and employment; the application of ethics frameworks, workplace diversity, social inclusivity and equality, reflection and dialogue, the language used to discuss AI, and the selection of applications and development paths.

Key messages

The report concludes with a short chapter that takes five recurring themes from the report and presents them as key messages.

Language matters. In many ways, the term 'AI' has become an obstacle to meaningful reflection and productive debate about the diverse range of technologies that it refers to. It could help to address the way we talk about AI – including how we identify, understand and discuss specific technologies, as well as how we articulate visions of what we really want from it.

Algorithms are subjective. Since human societies have structural biases and inequalities, ML tools inevitably learn these too. While the only definitive solution to the problem is to remove bias and inequality from society, AI can only offer limited support for that mission. However, it is important to ensure that AI counteracts, rather than reinforces inequalities.

AI is not an end in itself. The ultimate aim of supporting AI is not to maximise AI development per se, but to unlock some of the benefits that it promises to deliver. Instead of perfecting new technologies then searching for problems to which they could be a profitable solution, we could start by examining the problems we have and explore how AI could help us to find appropriate solutions. Meaningful dialogue with a range of stakeholders, including citizens, from the earliest stages of development could play a key role in defining what we aim to achieve, and how AI could help.

AI might fall short of its promises. Many AI applications could offer profound social value. However, employment impacts and privacy intrusions are increasingly tangible for citizens while the promised benefits to their health, wealth and environment remain intangible. The response could include targeting more ambitious outcomes while making more modest promises.

Europe needs to run its own AI race. AI is at a pivotal moment for both regulation and technology development and the choices we make now could shape European life for decades to come. In running its own race, European AI can ensure a meaningful role for citizens to articulate what they expect from AI development and what they are ready to offer in return, to foster a competitive market that includes European small and medium-sized enterprises (SMEs), and to put adequate safeguards in place to align AI with European values and EU law.

Glossary

The following glossary provides brief definitions of some key terms and concepts described in this study, along with their acronyms where they have been used.

5G	The fifth generation standard for communications technology, currently being rolled out, is anticipated to deliver higher speed internet and enable many more devices to be connected, see IoT.
AGI	Artificial general intelligence or 'strong AI' exhibits intelligence in a wide range of contexts and problem spaces, rather than only in specific niches.
AI	Systems that display intelligent behaviour by analysing their environment and taking action – with some degree of autonomy – to achieve specific goals.
Algorithm	A set of rules defining how to perform a task or solve a problem. In the context of AI, this usually refers to computer code defining how to process data.
Alife	Ideas and techniques are based upon fundamental biological processes, rather than intelligence or expertise.
ANN	Artificial neural networks process data to make decisions in a way that is inspired by the structure and functionality of the human brain.
API	Application programme interfaces are the access points that apps and third parties can use to engage with larger platforms and systems, like mobile phones or social media websites.
ASI	Artificial superintelligence refers to AI that that has higher levels of general intelligence (AGI) than typical humans.
Big data	Broader than AI, big data refers to sets of data that are so large and complex that they cannot be effectively stored or processed with traditional methods.
Black box	Systems, including several machine learning algorithms, whose operation is either inaccessible or too complex for the user to readily understand.
Data mining	Automated process for extracting data and identifying patterns and anomalies.
Data in the wild	Data that was produced and made available for one purpose and is then gathered and used for another purpose.
Error	In machine learning, the error is a measurement of distance between the algorithm's decision and the correct decision as understood through the labelled data or previous human decisions.
Explainability	In AI, explainability refers to how easily humans can understand and explain the inner working of algorithms, either for a specific decision or in terms of their overall logic.
Facial recognition	Includes facial verification (checking one face against data about one face, e.g. to check the identity of a phone user), identification (analysing many faces against data about many faces to identify individuals. Can also include facial classification to estimate age, gender or other categories such as mood or personality.

GAN	Generative adversarial networks are ANNs that learn by competing against each other, usually one producing content and the other detecting whether it was produced by an ANN or a human. They can produce realistic content as well as tools for detecting fake content.
GOFAI	Good old-fashioned artificial intelligence refers to first generation 'symbolic AI' systems that predate, and are often more explainable, than machine learning.
IoT	The internet of things is a system in which not only computers and smartphones, but many more devices such as home appliances, public, infrastructure as well as industrial and commercial tools, are connected online.
Labelled data	Data that is accompanied with information about the data, e.g. a picture of cat that is labelled as containing a cat.
ML	Machine learning refers to second generation AI techniques whereby the algorithm is designed to find its own solution to problems, rather than following rules defined by human experts.
Moore's law	The 1965 prediction that the number of transistors on a computer chip would double every two years. It still holds true, despite recently falling slightly behind schedule. Today, Moore's law is used more generally to refer to the trend for computer power to increase exponentially.
Narrow AI	Narrow or 'weak AI' refers to the current paradigm of AI tools which exhibit intelligence only in specific niches such as playing chess or recognising cats.
Reinforcement learning	A branch of ML where the algorithm develops policies for making sequences of decisions under different conditions through trial and error.
Singularity	The moment when AI becomes intelligent and autonomous enough to generate even more intelligent and autonomous AIs.
Strong AI	See AGI.
Supervised learning	In the context of ML, supervision refers to the use of labelled data to guide ML process.
Symbolic AI	First generation AI tools which, unlike ML, do not learn themselves. Their intelligence comes from programming human expertise directly into them.
Turing test	Turing tests define whether or not a machine is intelligent by examining whether they can be distinguished from humans under various conditions.
Unsupervised learning	Where no labelled data and minimum human intervention is used to guide the ML process. This could be by preference, or because appropriate labelled data is not available.
(AI) Winter	Periods in the past and, some fear, the future in which interest in AI is substantially reduced, accompanied by stagnation in investment, development and application.
Weak AI	See 'narrow AI'

Table of contents

1. What is artificial intelligence?

Artificial intelligence (AI) is probably the defining technology of the last decade, and perhaps also the next. This report provides an accessible review of how it works, why it matters and what we can do in response to the challenges it raises.

Since the earliest days of AI, its definition has focused on the ability to behave with the appearance of intelligence. Various forms of 'Turing test' declare machines as intelligent when humans cannot differentiate their actions from those of a human. Today's definitions of AI often include other requirements such as autonomy, and allow intelligence to be limited to specific domains. Rather than contributing to the proliferation of definitions,[1] this report adopts that of the 2018 European Commission Communication,[2] which is both accessible and typical of contemporary definitions:

> AI refers to systems that display intelligent behaviour by analysing their environment and taking action – with some degree of autonomy – to achieve specific goals.

This definition places no restrictions on the methods that are used to achieve intelligence. Indeed, AI is an umbrella term including a wide range of technologies and applications that have little more in common than their apparent intelligence, a quality which remains very much open to interpretation. Further, we regularly talk about AI that is already in widespread use alongside AI that is under development, and even AI that is speculated to possibly exist in the future. Consequently, the term 'AI' is regularly used to refer to any technique, used in any context – real or imagined – as long as it is somehow claimed to display features that some describe as intelligent. This inclusivity presents difficulties for assessing the impacts of AI development because, depending on which corner of the vast AI space is being considered, very different benefits and risks can be identified. As a result, AI is simultaneously high risk, low risk and everything in-between.

Since AI refers to so many technologies, applications and contexts, greater precision is required in order to hold a meaningful and constructive debate. For example, arguments about simple 'expert systems' used in advisory roles need to be distinguished those from those concerning complex data-driven algorithms that automatically implement decisions about individuals. Similarly, it is important to distinguish arguments about speculative future developments that may never occur from those about current AI that already affects society today.

The aim of this report is to support meaningful reflection and productive debate about AI by providing an accessible review of AI technology, impacts and options. The following chapter aims at demystifying AI, explaining its key approaches, how they work, their powers and limitations. The subsequent chapter examines the key opportunities and challenges presented by the application of these technologies. The penultimate chapter presents a range of options to respond to these opportunities and challenges via regulatory, technological and societal measures, and a concluding chapter presents some key messages.

[1] For a comprehensive review of definitions see Samoili, S. et al, Defining artificial intelligence, European Commission, 2020.

[2] Communication on artificial intelligence for Europe, COM(2018) 237, April 2018.

2. How does it work?

The following sections provide accessible introductions to some of the key technologies that come under the AI banner. They are grouped into three sections, which gives a sense of the chronology of the development of different approaches. The first wave describes early AI techniques, described as 'symbolic AI'. While these approaches can appear dated, they remain very relevant and are still successfully applied in several domains. The second wave describes more recent 'data-driven' approaches which have developed rapidly over the last two decades and are largely responsible for the current AI resurgence. The third section explores possible future waves of AI, focusing on approaches that remain far from the market. The aim is to equip readers with an understanding of key concepts and methods in AI, so they know what is 'deep' about deep learning and understand the difference between fuzzy logic and evolutionary methods.

2.1 First wave: symbolic artificial intelligence

Symbolic AI refers to approaches to developing intelligent machines by encoding the knowledge and experience of experts into sets of rules that can be executed by the machine. This AI is described as symbolic because it makes use of symbolic reasoning (e.g., if X=Y and Y=Z then X=Z) to represent and solve problems. This was the main approach to AI applications from the 1950s to the 1990s but, while other approaches dominate the field today, symbolic AI is still used in many contexts, from thermostats to advanced robotics. Here we describe two popular approaches within symbolic AI, expert systems and fuzzy logic.

2.1.1 Expert systems

In these systems, a human expert in the domain of the application creates precise rules that a computer can follow, step by step, to decide how to respond intelligently to a given situation. These rules, known as algorithms, are often expressed as code in an 'if-then-else' format. For example, to create a symbolic AI doctor, the human expert might start by writing the following pseudocode:

```
If the patient has fever
      Prescribe drug X
If the patient is coughing
      Prescribe drug Y
Else
      Send patient home
```

This is an example of **pseudocode**. It is not ready for a computer to read, but illustrates how an algorithm can work.

Symbolic AI can be said to 'keep the human in the loop' because the decision-making process is closely aligned to how human experts make decisions. Indeed, any intelligence in the system comes directly from human expertise being recorded in a 'machine readable' format that a computer can work with. Furthermore, humans can easily understand how these systems make specific decisions. They can easily identify mistakes or find opportunities to improve the programme, and update the code in response. For example, adding clauses to deal with special cases or to reflect new medical knowledge.

```
If the patient has fever and is allergic to drug X
      Prescribe drug Z
```

The example hints at the key drawback of this kind of expert system. In order to develop a useful and reliable system that works for complex and dynamic real-world problems, such as the work of a medical doctor, so many rules and exceptions would be required that the system would become

very large and complicated, very quickly. Symbolic AI is at its best in constrained environments which do not change much over time, where the rules are strict and the variables are unambiguous and quantifiable. Once such example is calculating tax liability. Tax experts and programmers can work together to develop expert systems that apply the rules that are in force for that tax year. When presented with data describing taxpayers' income and other relevant circumstances, the tool can calculate tax liability according to the rules and applying any applicable levies, allowances and exceptions.

2.1.2 Fuzzy logic: capturing intuitive expertise

In the expert system described above, each variable is either true or false. For it to work, the system needs to know an absolute answer to questions such as whether or not the patient has a fever. This could be reduced to a simple calculation of a temperature reading above 37 °C, but reality is not always so clear cut. Fuzzy logic is another approach to expert systems which allow variables to have a 'truth value' that is anywhere between 0 and 1, which captures the extent to which it fits a category. This allows patients to be assigned a rating of how well they fit the category of having fever. The figure might depend on the patient's temperature reading as well as other relevant factors such as their age or the time of day, and it allows the patient to be described as a borderline case.

This fuzzy logic is particularly useful for capturing intuitive knowledge, where experts make good decisions in the face of wide-ranging and uncertain variables that interact with each other. They have been used to develop control systems for cameras which automatically adjust their settings to suit the conditions, and for stock trading applications to establish rules for buying and selling under different market conditions. In each case, the fuzzy system continually assesses dozens of variables, follows rules designed by human experts to adjust truth values and uses them to automatically make decisions.

2.1.3 Good old-fashioned artificial intelligence

Symbolic AI systems require human experts to encode their knowledge in a way the computer can understand. This places significant constraints on their degree of autonomy. While they can perform tasks automatically, they can only do so in the ways in which they are instructed, and they can only be improved by direct human intervention. This makes symbolic AI less effective for complex problems where not only the variables change in real-time, but also the rules. Unfortunately, these are the problems where we need the most help. Millions of 'if-then-else' rules could not capture all of a doctor's domain knowledge and expertise, nor their continual development over time. Despite these limitations, symbolic AI remains far from obsolete. It is particularly useful in supporting humans working on repetitive problems in well-defined domains including machine control and decision support systems. The reliable performance of symbolic AI in these domains has earned it the endearing nickname 'good old-fashioned AI'.

2.2 Second wave: machine learning and data-driven artificial intelligence

Machine learning (ML) refers to a wide range of techniques which automate the learning process of algorithms. This differs from the first wave approaches whereby improvements in performance are only achieved by humans adjusting or adding to the expertise which is coded directly into the algorithm. While the concepts behind these approaches are just as old as symbolic AI, they were not applied extensively until after the turn of the century when they inspired the current resurgence of the field. In ML, the algorithm usually improves by training itself on data. For this reason, we talk

about data-driven AI. Practical applications of these approaches have really taken off over the last decade. While the methods themselves are not particularly new, the key factor in recent advances in ML is the massive increase in the availability of data. The tremendous growth of data-driven AI is, itself, data-driven.

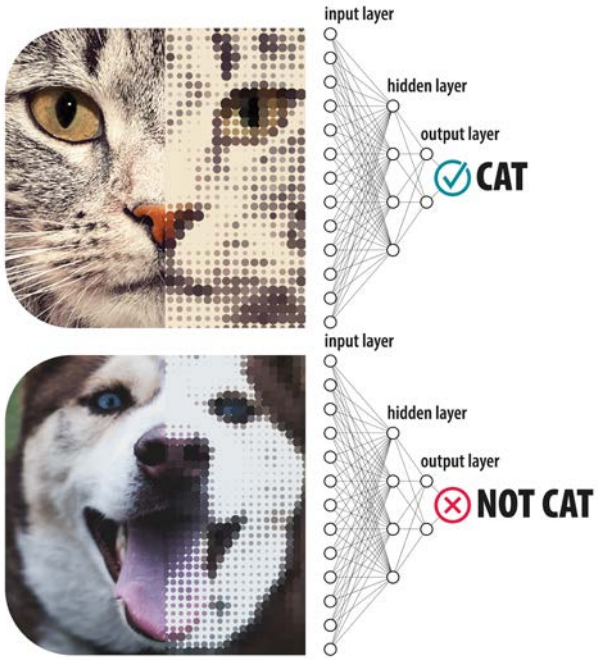
Usually, ML algorithms find their own ways of identifying patterns, and apply what they learn to make statements about data. Different approaches to ML are suited to different tasks and situations, and have different implications. The following sections present an accessible introduction to key ML techniques. The first provides an explanation of deep learning and how software can be trained before exploring several concepts related to data and the important role of human engineers in designing and fine-tuning ML systems. The final sections illustrate how ML algorithms are used to make sense of the world and even to produce language, images and sounds.

2.2.1 Artificial neural networks and deep learning

As the name suggests, artificial neural networks (ANNs) are inspired by the functionality of the electro-chemical neural networks found in human (and other animal) brains. The working of the brain remains somewhat mysterious, although it has long been known that signals are transmitted through a complex network of neurons and, in doing so, both the signal and the structure of the network are transformed. In ANNs, inputs are translated into signals that are passed through a network of artificial neurons to generate outputs that can be interpreted as responses to the original inputs. The learning process refers to the transformation of the network so that these outputs are useful – or intelligent – responses to the inputs.

ANNs process data that is sent to the 'input layer', and generates a response at the 'output layer'. In between, there are one or more 'hidden layers', which manipulate the signals as they pass through them. The basic structure of an ANN is shown in Figure 1, with an illustrative example of an ANN that can predict whether or not an image depicts a cat. First, the picture is split into individual pixels which are sent to the neurons in the input layer of the ANN. From there, they are sent as a signal to the first hidden layer. Each neuron in this hidden layer receives several signals, which they combine and manipulate to generate a single output signal. While Figure 1 shows only one hidden layer, ANNs usually contain several sequential hidden layers. In those cases, this step is repeated with signals passing through each hidden layer until they reach the final output layer. The signal generated at the output layer is the final output, which is interpreted as a decision about whether or not the image depicts a cat.

Figure 1 – Schematic of an artificial neural network for recognising images of cats



Graphic by EPRS, produced by Samy Chahri; picture credits: @bedneyimages (freepik.com) and C. Brear (Unsplash).

Now we have a simple ANN, inspired by a simplified model of the brain, which can respond to a specific input with a specific output. The ANN is not really aware of what it is doing, or even what a cat is, but if we give it a picture, it will always tell us whether or not it 'thinks' it contains a cat. The question is, how can we develop an ANN that gets the answer right? First, it needs to have the right structure. For simple tasks, ANNs can work well with just a dozen neurons in a single hidden layer. Adding more neurons and layers allow ANNs to tackle more complex problems.[3] Deep learning simply refers to ANNs with at least two hidden layers, each containing many neurons. Having more layers allows ANNs to develop more abstract conceptualisations of problems by splitting them into smaller sub-problems, and to deliver more nuanced responses. While in theory three hidden layers may be enough to solve any kind of problem, in practice ANNs tend to contain many more. For example, Google's image classifiers use up to 30 hidden layers. The first layers search for lines they can identify as edges or corners, the middle layers try to identify shapes in these lines, and the final layers assemble these shapes to interpret the image.

Virtual machines
Computers execute code sequentially, line-by-line, as illustrated in the expert system described above. In ANNs, however, signals seem to pass through simultaneously as the signals are processed in parallel. In fact, the ANN is a 'virtual machine', which has to be translated into a sequence of commands before it can be executed on a physical machine. The output is the same as it would have been if it were processed in parallel.

So, if the 'deep' part of deep learning is about the complexity of the ANN, what about the 'learning' part? Once the right structure of the ANN is in place, it needs to be trained. While in theory this can be done by hand, it would require a human expert to painstakingly adjust neurons to reflect their own expertise of how to identify cats. Instead, a ML algorithm is applied to automate the process. In the following sections, two significant ML techniques are explained. The first applies calculus to make gradual improvements to individual ANNs, while the second applies evolutionary principles to yield gradual improvements in large populations of ANNs.

2.2.2 Training neural networks: back propagation and gradient descent

If we compare the actual output of an ANN to the desired output as reported in the labelled data, the difference between the two is described as the error. ML algorithms such as back propagation and gradient descent aim to gradually improve the ANN's performance by minimising this error. They do this by adjusting the ANN and checking whether the error has reduced before re-adjusting. This process is best explained through calculus, however the following paragraphs provide an accessible introduction.

Labelled data and supervised learning
Supervised learning refers to the use of labelled data – such as pictures that say whether or not they contain cats – to train algorithms. These approaches devise their own methods of predicting how pictures should be labelled. Unsupervised learning may be used where good quality labelled data is not available. They excel in finding new clusters and associations within data which might not otherwise have been identified or labelled by humans. Since labels are often incomplete or inaccurate, many applications such as content recommendation systems combine both supervised and unsupervised ML approaches.

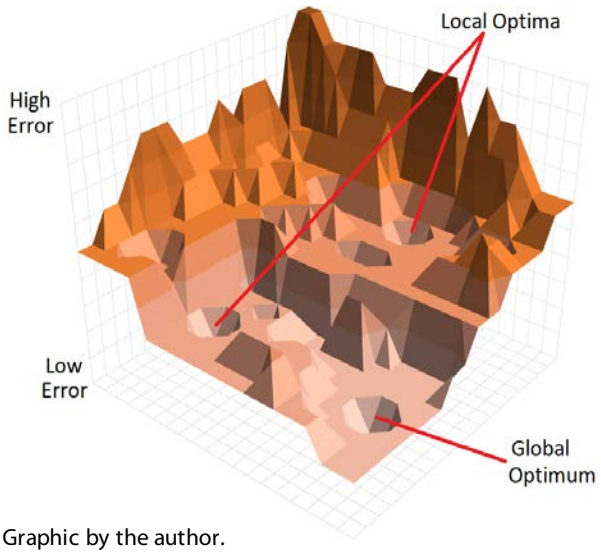
Back propagation deals with adjusting the neurons in the ANN. The process starts with the routine described in the previous section, where

[3] Human brains contain ~100 billion neurons;, cockroaches ~1 million; and snails ~10 thousand.

an input signal is sent to the ANN, passes through the hidden layer(s) to the output layer, and generates an output signal. The error is then calculated by comparing the output to what it should have been according to the labelled data. Now, the neurons are changed to reduce the error, so the ANNs output is more accurate. This correction process starts with the output layer, which has a stronger influence over the results, and then makes changes further back through the hidden layer(s). It is called back propagation because the correction of the error propagates backwards through the ANN.

In theory, it is possible to calculate the error for every possible ANN. That is, to generate a set of ANNs with every possible combination of neuron, test each of them against the labelled data, and choose the one with the lowest error. In practice, however, there are too many possible configurations for this to be feasible. The AI engineer needs to find a way of being more selective with a smarter search for the lowest error. This is where gradient descent comes in. Imagine a chart of every possible ANN, each point representing one ANN, with the altitude representing its error. This would form an 'error landscape', as illustrated in Figure 2.[4] Gradient descent is a method of trying to find the lowest point on this landscape – the ANN with the lowest error – without having access to the map.

Figure 2 – An error landscape



Graphic by the author.

Gradient descent is often compared to a hiker that needs to find their way down a mountain, but it is so foggy that they can only see one metre in each direction so they adopt a strategy of looking around, deciding which direction offers the steepest descent, moving in that direction, and then looking around again and repeating the process until they find their way down the mountain. Similarly, an ANN can be generated, located at a random point on the error landscape. Its error is calculated, as well as the error resulting from a few different kinds of adjustment which correspond to nearby positions on the error landscape. The adjustment that offers the best improvement is assumed to be the best direction, so the changes are implemented and then the process is repeated with a new set of tests. Just as the hiker takes the steepest possible immediate descent that they can see in their immediate vicinity, the ANN makes gradual improvements until it reaches the best solution that it can find.

While the algorithm might find the best possible solution – the 'global optimum' – the approach is not perfect. Just as we might expect the unfortunate hiker using this strategy to get stuck in a recess before descending the mountain – their dogmatic refusal to ascend a little limiting them from climbing out of a ditch – the algorithm can settle for a 'local optimum' that is not the best solution

[4] Here, the possible ANNs are mapped across two dimensions, with error as the third dimension, so that it is easy to visualise. This would be accurate if there were only two variables in the ANN but, in reality, they have so many neurons that the landscape would have millions of dimensions, making it too difficult to calculate and impossible to represent on paper.

available, but the minor modifications would make the situation worse before it could get better. This is why, in practice, the whole exercise is repeated many times, starting from different points and using different training data.

Gradient descent and back propagation make use of labelled data to calculate the error. If all the data was used for the learning process, however, the algorithm might memorise the training data without having gained any useful capacity to respond to new data. In order to ensure this does not happen, some of the labelled data is not used in training, and are used only to test the results. But what if there is no labelled data at all? In the following section, an alternative ML approach is presented that sidesteps the problem by evaluating more general behaviour across populations of ANNs.

2.2.3 Training methods inspired by nature

While gradient descent and back propagation are based upon mathematical concepts such as calculus, another set of methods are inspired by evolutionary concepts such as survival of the fittest, reproduction and mutation. There are many approaches within this family of evolutionary training methods, but the broad principles remain the same. A population of ANNs is created. They compete against each other and are subjected to artificial selection – the AI equivalent of natural selection – so that those that perform badly are filtered out, while those that perform well survive to the next generation. To replenish the population, new ANNs are generated through AI's answer to reproduction. These could involve the combination of different aspects of one, two, or any number of parent ANNs, along with a dose of random mutation.

Let's explore how an evolutionary approach can be applied to train ANNs to play chess. We could start by creating a population of 100 random ANN 'players', and make them play against each other, so they take turns in receiving inputs describing the position of the pieces, and generate outputs that are interpreted as moves. This first generation of untrained players will not be very good at the game, but some will inevitably be 'less bad' than others and win some games. At this point, the ANNs can be ranked. In keeping with the principle of the survival of the fittest, the worst players are deleted. Better players survive, and their 'genetic material' – in the form of layers of ANNs – are combined and mutated to produce a new generation of ANNs which join them in the next round of games. The new ANNs will respond differently to signals, so some may play better than their 'parents' and others worse.

Since only the better players survive and shape the features of future ANNs, the environment is conducive to a steady overall improvement as the generations pass. The approach can even yield champion players that beat most humans. The interesting thing about evolutionary methods is that they yield results without any human expertise on the problem, without labelled data from previous games, without even having access to the rules. However, those that tend to play well, tend to survive. This means ANNs can develop interesting ways of deciding how to play the game well, including strategies that humans have never thought of, and may have trouble appreciating.[5] If an engineer is asked to explain why such an ANN made a move then they might show how its response was mathematically determined by its structure, but they cannot always explain why this structure generates good moves. This leads to the problem of the transparency of algorithms, which will be discussed later.

[5] This is well illustrated in the documentary film *AlphaGo*, Moxie Pictures, 2017.

Evolutionary approaches can be applied to other kinds of optimisation problems such as improving computer programs or transport schedules. There are also many other interesting AI techniques inspired by biological and behavioural mechanisms observed in nature. For example, ant colony optimisation is modelled on how ants use pheromones as signals to find and highlight the quickest route between two locations, and can be used to optimise vehicle navigation and telecommunication networks. Hunting search is a search and optimisation technique based upon pack hunting by lions, wolves and dolphins. 'Swarm intelligence' techniques inspired by the honey bee's dance (among other apian behaviours) have been applied in modelling and optimisation tasks in electrical, mechanical, civil and software engineering.

Reinforcement learning (RL) is another branch of ML which focuses on developing a policy for making sequences of decisions under different conditions. It is particularly useful when a system could be subject to a wide range of conditions, each with different implications for appropriate action. First the RL algorithm identifies some features of the conditions and attempts some actions, it then receives feedback about the quality of the response, which is used to maintain a set of scores for different combinations of conditions and actions. RL is sometimes compared to how children learn to walk, and its trial and error techniques have been deployed for training self-driving cars. In many respects, the process is similar to back propagation and gradient descent but, while those methods require large amounts of labelled data prepared in advance, RL allows constant adjustment of behaviour to learn in real-time.

Gradient descent and evolutionary methods can be intensive and complicated to set up but, once the training is complete, the resulting ANN can be used to process new data very quickly and efficiently. In reality, however, algorithms are constantly updated in response to new data availability, bugs that are identified, developments in techniques and changes to the problem that the algorithm seeks to solve. Also, several of these methods are often combined in methods (such as 'deep reinforcement learning' which integrates elements of deep learning and reinforcement learning) and applications (game playing AI often synthesises elements of symbolic AI and ML).

2.2.4 Data mining, big data and data in the wild

Since data is so central to contemporary AI development, several data-related concepts are frequently raised during debates about AI. AI engineers spend as much time thinking about data as they do about algorithms. They need lots of good quality data to perform effective ML and test the results. 'Data mining', is a field of computation focused on the automated identification of patterns and anomalies in data sets. The data set could be anything from measurements of underground geological formations to text found on social media, and the mining process could deploy ANNs, statistics and modelling to identify useful features. 'Big data' refers to datasets that are so large and complex – including content from different sources, in different formats, and with different degrees of authenticity and accuracy – that they cannot be stored or processed in the same way as smaller datasets. This brings us to 'data in the wild', usually referring to data that was produced for one purpose but remains somehow accessible and is gathered or used for some other purpose. Depending on the circumstances, use of this data can be unreliable, unethical, and even illegal.

2.2.5 The art of artificial intelligence

It might be tempting to think of ML as doing all the hard work, but the ML algorithm can only follow the precise instructions of its creator. This section highlights the art of the AI engineer. They harness the power of concepts from a range of disciplines – most notably computing, logic, statistics and

calculus – while balancing a range of considerations about the problem itself and the context of its solution.

First, the engineer needs to find a good way of encoding the problem itself. For the chess-playing ANN, the engineer needs to express the positions on the board as a signal to be sent to the input layer. They also need to find a way of interpreting the output as a valid move. This means either designing the output layer so that its signal can always be interpreted as a legitimate move, or devising a strategy for managing any illegitimate moves suggested by the ANN.

If the ML algorithm uses training data, the AI engineer must consider which data to use and how. Where 'data in the wild' is used, they must ensure that it is legal and ethical. Even inadvertent storage and processing of some content – such as terrorist propaganda and child pornography – can be illegal. Other data might be subject to copyright, or require 'informed consent' from the owner before it is used for research or other purposes. If the data passes these tests, the engineer must determine whether it is sufficiently large and representative to be suitable for the problem at hand. A dataset for learning to recognise cats should contain lots of pictures from different angles, of different colours, and any labels should be accurate. Finally, the engineer needs to decide how much data to use for training, and how much to set aside for testing. If the training dataset is too small, the ANN can memorise it without learning general rules, so they perform poorly when tested with new data. If the testing dataset is small, there is less scope to evaluate the quality of the algorithm.

The AI engineer also needs to make several important decisions about the structure of the ANN and the ML algorithm. The ANN needs enough neurons and layers to deal with the complexity of the problem. Too few and the ANN will not be able to deal with complex problems, too many and they tend to memorise the training dataset instead of learning general rules. For gradient descent, they need to define how many evaluations to make before deciding on a direction to travel, as well as how far to travel in the chosen direction before re-evaluating. This is known as the 'learning rate'. If it is slower, the algorithm takes more time but makes better choices, like the lost hiker taking small careful steps. If it is faster, it adapts more quickly but might miss important features, rather as though the hiker runs blindly through the fog. The engineer must consider the problem and decide how to balance speed against accuracy.

In evolutionary approaches, the AI engineer has to decide the population size and number of games to play, balancing thorough evaluation against processing burden. They also need to decide how many ANNs to delete per generation, and how combination and mutation are used to create new generations. Mutation is important for the emergence of new solutions, but if it is too strong then the 'offspring' might be so different from their parents that they perform as badly as the randomly generated ANNs from the first generation of the process.

A further question is raised in deciding when a solution of sufficient quality has been found. As discussed in the context of gradient descent, an algorithm can get stuck in a 'local optimum', which is definitely the best solution in the vicinity, but not necessarily the best possible solution. Similarly, evolutionary populations can develop into a local optimum, whereby the parent ANNs cannot produce offspring that perform better than them, even though better solutions are available. The engineer can counteract local optima by adjusting the learning rate in gradient descent, or altering the approach to reproduction and mutation in evolutionary methods. They can also repeat the training process several times with different starting points and data sets. This is often worth the effort because, while ANNs are difficult to train, once a good solution is found they can be applied to new data very quickly.

Many of these decisions require the AI engineer to balance the constraints of the problem at hand, its context, and the data and processing resources available to them. There are no objectively correct formulas, so these decisions are part of the 'craft' of AI. Practitioners rely upon experience, intuition, experimentation and shared wisdom to make effective decisions. Perhaps unsurprisingly, decisions about the structure of the ANN, the mutation rate, learning rate and population size are also sometimes delegated to ML. Nonetheless, the AI engineer still needs to make difficult decisions and give precise instructions about when and how much the ML can adjust each variable. These trends take the AI engineers even further away from the subject expertise embedded in their algorithms. Nonetheless, the engineer retains an irreplaceable role in the design and optimisation of the environment in which machines learn.

2.2.6 Making sense of the world: identifying language, images and sounds

This section focuses on how second wave AI algorithms are deployed to make sense of the world. That is, how they can respond in useful ways to language, images or sounds. A good example of this is spam detection. When users identify emails as 'spam', they provide labelled data that is used to train an ANN to identify emails that look like spam, which are then automatically filtered into a junk folder. In 2015, Google reported[6] that its Gmail service mislabelled .05 % of wanted emails as spam. For an average user that's around two emails per month, so it's far from perfect, but each time users label an email in their junk folder as 'not spam' they provide more labelled data to train and improve the ANN. Major platforms have access to more data, so they can develop more accurate tools. Since this in turn attracts more users, cycles of market dominance can emerge that will be further discussed in the next chapter.

First wave symbolic AI translation tools tried to encode the expertise of translators into rules for converting text from one language to another, but the approach largely failed as the rules proved to be too cumbersome. Early data-driven solutions bypassed these rules by finding clusters of words in sources that were already translated by experts – notably including documents translated by the European Parliament – and stitching them together. Today, top translation tools use the same data to train ANNs to find their own rules for translating text directly, even if the specific combination of words is not present in the corpus. Unlike symbolic AI approaches, the developers of ML translation do not need to speak the languages, or even study their grammar.

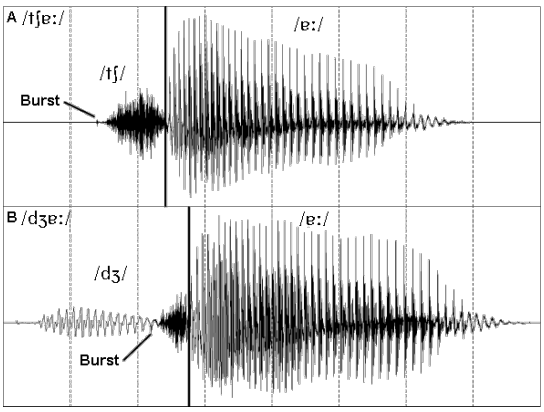
Sound and vision are more complicated than text. Even as recently as 2005 – during the author's training as an AI programmer – image and speech recognition were taught in the symbolic AI tradition, as problems of encoding human expertise. For example, to recognise pictures of cats, the algorithm would search pixel-by-pixel for lines that look like edges and enhance them to identify outlines which they compare to templates corresponding to cats. Similarly, algorithms for speech recognition followed painstakingly hand-programmed instructions to find patterns in sound waves when people speak and identify them as 'phonetic units' (the basic range of speech sounds, see Figure 3) before translating them into meaningful combinations of letters and, eventually, words.

These expert-trained image and speech recognition systems were difficult to develop, and struggled with new images and voices. They have now been largely replaced by data-driven ML tools. Speech to text tools are now trained on labelled data, such as transcripts of audio files. Image recognition tools are trained on millions of pictures from the internet that are already labelled. Indeed, Facebook

[6] See: The mail you want, not the spam you don't, the official Gmail blog, Google, 2015.

trained its face recognition service on millions of photos that were diligently labelled by its users. The approach is not infallible, and depends to a large extent on the quality of the data. In one well-known example, an algorithm trained to differentiate dogs from wolves performed well against the data that was used, but worked by classifying the canines by the presence or absence of snow in the picture.

Figure 3 – Identification of phonetic units through waveform analysis of sounds tʃɐ (as in chat) and dʒɐ (as in 'jam').



Source: © Macquarie University

We can examine what is going on 'under the hood' of image recognition ANNs by using them to enhance the features of an image that resemble the object it was trained to recognise. Starting with a static image and repeating this recognition and enhancement process, the ANN gradually reveals, in abstract and somewhat blurry images, its 'understanding' of what the object looks like. This kind of feature extraction and enhancement has yielded interesting results. The example shown in Figure 4 below presents ethereal images generated by an ANN that was trained to recognise dumbbells. Some say the example is evidence of a kind of autonomous creativity. More likely, the psychedelic images are the AI equivalent of human pareidolia, where people identify patterns that they are trained to recognise – notably faces – in clouds and other objects. Here, it is particularly interesting to note how the ANN's ideal of a dumbbell always features a human arm. In this case, it seems the training data contained many images of dumbbells in use, but not enough labelled images of dumbbells on their own.

Figure 4 – Images of dumbbells, featuring arms, produced by an ANN trained on images of dumbbells

Source: Google Inc (CC by 4.0)

These tools are good at individual patterns, and can carry out some useful tasks with visual and linguistic material, but do they really understand the world? To answer this question, first, we can consider word vectors, which describe the 'proximity' of words to each other. Proximity is calculated by how often, statistically, a word is used in the same context as another. So, for example, the word 'king' has a closer proximity to 'man' than it does to 'woman'. The approach enables impressive linguistic manoeuvres such as 'king - man + woman = queen'. It is certainly a well-designed technique, and may have some useful applications, but the meaning of words cannot be reduced to their statistical occurrence. The tool does not know what a monarchy is and has no understanding of why such concepts matter. While it can convincingly use words that relate to each other such as 'king' and 'queen' or 'patient' and 'cancer', it cannot appreciate the human sentiments associated with these things because their experience of them is limited to the statistical occurrence of words that humans use to represent them. This is important to appreciate because, in some application areas, it might be dangerous to believe that AI really understands something when really it is just very good at behaving as though it does.

2.2.7 Imagination and creativity: producing language, images and sounds

This section focuses on how AI can be deployed to create language, images or sounds. In one sense, all ML algorithms are creative inasmuch as they create their own ways of solving problems without expert advice. In another sense, algorithms can only follow precise sequences of instructions, so it is difficult to describe them as imaginative. Illustratively, computers cannot even spontaneously imagine the random numbers they need to simulate dice rolls or generate ANNs. The best they can do is follow precise instructions to produce numbers that are so unpredictable and well distributed that they appear to be random. For example, precise measurements of the milliseconds between two keystrokes can be taken, with the figures to the right of the decimal place used as though they were a sequence of random numbers. Measurements of sufficiently random external processes can also be used, such as radioactive decay, atmospheric noise and even images of lava lamps. Like AI tools, the methods are interesting and useful but the impression of understanding or creativity fades as the mechanisms are exposed as strictly procedural. For AI, it is important to recognise these limits, in particular the difference between 'understanding and creating' and 'appearing to understand and create'.

Many articles are now written by AI 'journalists' rather than the human equivalent. They take basic information in a set format – perhaps weather predictions, sports results or stock market performance – and follow rules to identify the most important information before converting it into sentences that read like human written news articles. Here, the algorithm needs to ensure that the text maintains a consistent style while minimising repetition of words and sentence structures that would seem too robotic to the reader. Once an article is written – by an AI or human journalist – AI subeditors can be deployed to generate headlines and automatically test their attractiveness to users and search engines before settling upon a final choice.

Basic language production is widely offered on smartphones. This includes predictive text to support typing, 'suggested replies' to bypass typing, and virtual assistants which produce spoken text to interact with the user. Predictive text is quite simple, linking each word to others that commonly follow them, much like the word vectors described above. These start with generic lists, but can be quickly customised to follow an individual user's style. So, when typing 'how are', the AI would suggest 'you?'. AI music generators are similar, they learn to mimic the composition of music by guessing what kind of sound a human musician might create next. Suggested replies are similar, but operate on whole messages rather than individual words. They first capture the important points of an email received, such as an invitation to meet for coffee, and suggest a few responses based upon the style and type of responses typically given to that kind of message such as 'sure' and 'sorry, I can't'.

Virtual assistants (VAs) are even more complex, as they start with the sound of the user's speech when they ask to play a song, send a message or buy a product. These sounds need to be converted into phonetic units, then into words, and then into a set of instructions to follow – perhaps requiring access to the user's diary, contacts, location and a myriad other accounts and services. Similarly, the VA's requests for further information need to be converted into words and synthesised into speech. Just as the VA needs to take account of the user's voice and accent, the synthesis of their own speech also depends on the user's preferences for voice and accent. Furthermore, just as good AI journalists avoid producing dull and repetitive text, a good VA will use some melody, and diverse expressions. Some even add interjections like 'hmmm' while pretending to think about a response and give emotional cues that are appropriate to the content. All of this effort is made to make them seem more human to the user. As more people use these services, more data is collected to train the VA to

identify their users' voices, accents, emotions, interjections and – perhaps most importantly – their shopping habits. However, it is important to recognise that while the VA's calculated gestures may be useful and appear convivial, they do not represent deep understanding of concepts, nor empathy for users.

Imagine asking a VA to book a table for dinner. It calls the restaurant and generates human-like speech to request a reservation. The restaurant's VA answers the call, generating its own human-like speech to discuss availability. The two AI agents interact in spoken English, complete with interjections, hesitations and emotional cues. Furthermore, each records the speech of the other and uses it to train themselves to speak 'more humanly' next time. If they were left to train on each other like this, the VAs could eventually develop their own accents or even dialects.

Figure 5 – ANNs can turn audio into realistic lip-synched videos.

Source: University of Washington

Indeed, this kind of interaction between different ANNs is precisely how some newer ML algorithms known as generative adversarial nets (GANs) work. GANs can generate realistic images by training a 'detective ANN' to recognise whether a picture was produced by a human or a computer, then training a 'forger ANN' to produce images, which are tested by the detective ANN. In the process, the pair of ANNs both get better, the detective at identifying fake images and the forger at producing realistic images. These forger ANNs can be used to develop interesting image modification and production tools. They can adapt pictures to make people appear older or younger, and generate convincing 'new faces'. The same principles can be used to develop tools for producing realistic sounds and videos. Indeed, such techniques have led to the emergence of tools for producing extremely realistic videos known as 'deepfakes', which have been used to create fake pornographic videos featuring celebrities, and fake statements featuring politicians (see Figure 5). Of course, the falsification of images is not new, but deepfakes can be incredibly realistic and are, perhaps, easier to tailor and mobilise for specific objectives. Depending on the jurisdiction, some deepfakes may be illegal, while others are considered as legitimate satire, critique or creative reuse.

2.3 Speculative future waves: towards artificial superintelligence?

The approaches set out in the previous sections are described as 'weak' or 'narrow' AI, in the sense that they can behave intelligently in domain-specific niches such as chess or cat recognition. 'Strong' or 'general' AI (AGI), on the other hand, is closer to our understanding of human intelligence as it refers to algorithms that can exhibit intelligence in a wide range of contexts and problem spaces. If weak AI is already rather strong, AGI would provide a new paradigm of capability. However, since it does not yet exist, it belongs to the realm of speculative AI. A second key term from the speculative domain is artificial superintelligence (ASI),[7] that is, with higher levels of general intelligence than typical humans. A third is singularity, which refers to the moment where AI becomes sufficiently

[7] Biological superintelligence – achieved through real brains – is not explored here. However, the possibility has been considered through natural (or artificial) selection, genetic engineering, better organisation of collective human intelligence, and the use of drugs or implants to enhance cognitive function.

intelligent and autonomous to generate even more intelligent and autonomous AIs, breaking free from human control and embarking on a process of runaway development.

There is some debate about whether these speculative AIs could be achieved by incremental development of existing technologies and techniques. Some experts cite Moore's law of continual exponential advancement in computer power, or suggest that today's AI could be deployed to produce the next generation of AI. However, most experts agree that there are fundamental limits to both Moore's law and the capabilities of the current AI paradigm. Amongst these critical thinkers, some argue that paradigm-shifting development could eventually make AGI possible – and perhaps even inevitable – while others are more sceptical.

These final sections of this chapter present several ideas about possible future waves of AI and how they could be achieved. Some of them are radically new, while others combine or extend elements of current approaches. Some are proposed as routes to ASI, while others have more modest and practical aims. Some are subject to active research and development, while others are more akin to thought experiments. They all, however, face substantial technical barriers to implementation and, as such, they all remain speculative future possibilities with no guarantee of realisation.

2.3.1 Self-explanatory and contextual artificial intelligence

The next wave of AI might combine aspects of the first and second waves of AI in ways that deliver completely new functionality.[8] They could maintain the complex and sophisticated power associated with ANNs while also delivering the same level of explainability associated with expert systems. These AI agents could also be capable of taking account of wider contextual knowledge about the world in order to provide more accurate results from smaller sets of training data. For example, such a tool for recognising animals might be trained with just two or three photos of an animal it has never seen before, but be capable of drawing on its wider knowledge of the world – e.g. about animal movements, bodies, and how they might look from other angles or positions that they have not seen before – to reliably identify the animal in other images. Similarly, a handwriting recognition system which is trained on images of written text might also draw on their contextual knowledge of how people use pens to write in order to help them to decipher what is written. The aims of self-explanatory and contextual AI may appear modest (e.g. in comparison to the developments speculated below) but the potential of – and barriers to – adding these features to today's AI should not be underestimated.

2.3.2 Robotic artificial intelligence

If AI is the brains, then robotics is the brawn. In common with AI, robots can undertake some tasks with superhuman ability (e.g. lifting heavy weights), while failing spectacularly at things that most humans find easy (e.g. walking up staircases). Robotics is not really a path to AI in itself, but is a complementary field with potential synergies that could lead to substantial advancement in certain application areas. For example, ML can be applied to help robots to manipulate physical objects with greater autonomy, flexibility and dexterity, which could make automated production and distribution more efficient. AI and robotics could be combined to design, build and control new computer hardware and robotics which strengthen both fields, initiating a cycle of ongoing synergetic development and advancement.

[8] See, for example, work at the United States Defense Advanced Research Projects Agency (DARPA), 2017.

Here, it is also worth mentioning that the marriage of AI and robotics is a major area of development for military technologies, notably in autonomous weapons systems. At present, drones are remotely piloted by humans, but this introduces several weaknesses, including communication channels that are vulnerable to detection and attack, as well as much slower human decision and response times than with automated control systems. Full AI command resolves both issues while opening new opportunities such as swarming capabilities which are beyond human capability. Such systems are not beyond today's technical capabilities, but the field is controversial and 'human-in-the-loop' policies dominate, as discussed in the next chapter.

2.3.3 Quantum artificial intelligence

Quantum computers harness the power of simultaneity to quickly find solutions to very complex problems, promising a revolutionary increase in computing power. If the problem is to find a one-in-a-trillion combination that works as a solution, a normal computer would have to check each possibility one by one, while a quantum computer can check them all at the same time, in a single operation. This means they are particularly well-suited to problems such as simulating environments, finding solutions, and optimising them. Since these kinds of problems are central to AI, developments in quantum computing could enable significant advances in the field.

What is quantum computing?
Single bits of data on normal computers exist in a single state, either 0 or 1. Single bits in a quantum computer, known as 'qubits' can exist in both states at the same time. If each qubit can simultaneously be both 0 and 1, then four qubits together could simultaneously be in 16 different states (0000, 0001, 0010, etc.). Small increases to the number of qubits lead to massive increases (2^n) in the number of simultaneous states. So 50 qubits together can be in over a trillion different states at the same time. Quantum computing works by harnessing this simultaneity to find solutions to complex problems very quickly.

While there have been some promising recent breakthroughs in quantum computing, their details often serve to illustrate how far the technology is from launching on the market. For example, in late 2017, IBM's 50-qubit machine broke industry records by remaining stable for 0.00009 seconds. Two years later, Google claimed quantum supremacy when their 54-qubit machine completed a calculation in 200 seconds that might have taken a non-quantum supercomputer up to 10 000 years to complete. However, while the machine is an impressive proof of concept, it is not yet capable of performing calculations with specific practical uses. A general-purpose quantum computer would require closer to 1 million qubits operating near absolute zero (- 273 °C). As such, it seems that reliable and useful quantum computers will probably remain unavailable for at least the next decade. Some suggest it is a moving target that will always remain tantalisingly out of reach. Here, it is enough to note that quantum computing is a speculative development that, if achieved, could enable the emergence of future waves of AI either by applying current methods more effectively, or by enabling the development of completely new approaches.

2.3.4 Evolving artificial superintelligence

One suggested path to ASI is to develop increasingly sophisticated ANNs through better evolutionary methods running on more powerful computers. Evolving superintelligence could start with the design of an algorithm to generate huge populations of multiple species of ANN in an immense simulated evolutionary environment. It took millions of years from the emergence of the first biological neurons to the evolution of intelligent humans and, during this time, a huge range of

lifeforms occupied the most sophisticated and complex environment known to humanity: Earth. Current processing capabilities could not do justice to evolutionary environments of that scale.

However, some shortcuts to superintelligence may be possible in computer environments. Biological evolution to date was engaged not only with intelligence, but also the development of complex organs and physical defences, capabilities such as flight, and features such as synergetic co-dependence between bacteria and other organisms. Biological development can also get stuck in unproductive evolutionary ruts, much like ML algorithms get stuck in local optima. An AI simulator could skip these and many other time-consuming biological processes such as maturing and ageing, pluck emerging populations out of evolutionary dead-ends and sidestep the unnecessary distractions associated with bodily survival and reproduction. By selecting exclusively for general intelligence, perhaps it could develop more quickly for computers than it did for humanity. Could this lead to the emergence of AGI and, by running the simulation even longer, towards the emergence of ASI?

The answer is not clear. Because of the demands of the environment in which they evolved, humans became good at recognising animals and understanding their movements, but not at making quick and complex mathematical calculations. Likewise, the kind of capabilities the simulated populations develop would depend on what kind of challenges they face and the resources they can draw upon to develop solutions. Since they would not have human-like bodies or biological needs, they would be unlikely to develop human-like languages or societies. Given this, they might never face the human-like problems that would drive them to develop human-like solutions. This does not mean they could not develop nuanced and surprising solutions to interesting problems, it is more a question of whether humans could relate to these problems and solutions in a useful or meaningful way. For this reason, human intelligence seems to require immersion in human society which, in turn, seems to require the presence of human bodies.[9]

2.3.5 Brain emulation and artificial consciousness

Another proposed approach to developing AGI would maintain alignment with human intelligence by producing a very detailed digital copy of the human brain, including all of the neurons and their connections of various strengths. If we speculate that we might have the technical capacity to do this, and a sufficiently accurate and complete understanding of the brain, the result might be a complete digital emulation of a mind, with the capacity to process sensory inputs, to remember, to learn, and to apply general intelligence. Perhaps the steps to developing an ASI would be as simple as speeding up the computer or running it without degradation for centuries, so that it has the time to carefully study every subject in every language. It could even be enhanced with modules allowing them to perform advanced mathematical calculations or to access the internet directly.

Since the ANN would need to simulate around 86 billion neurons and about 150 trillion connections in real-time, full brain emulation remains firmly in the domain of speculation. Nonetheless, serious projects are pursuing the ambition, including the billion-euro EU-funded Human Brain Project, which has made some progress in mapping mouse brains, although the models are incomplete and operate much slower than real time. It is not clear whether such an emulated brain would need sleep, how limited their ultimate capacity for memory or knowledge might be, nor whether they would experience pain, sadness, existential terror or consciousness.

[9] H. Collins, *Artifictional intelligence: Against humanity's surrender to computers*, Polity Press, 2018.

2.3.6 Wetware and biological systems

The nascent field of artificial life (Alife) differs from AI in that its ideas and techniques are based upon fundamental biological processes, rather than intelligence or expertise. Nonetheless, it does have some crossovers with AI, particularly in the context of evolutionary learning approaches and other methods inspired by nature. Much like AI, Alife is primarily developed via software (computer code and data) and hardware (physical components), but it can also involve 'wetware', which refers to the use of biological materials as components of the system being developed. In other fields, such as gene editing (the modification of DNA) and synthetic biology (the creation of artificial biological systems), wetware takes centre stage. Both synthetic biology and gene editing may benefit from AI insights, and they are identified as a possible pathway towards some future bio-AI, although this remains a distant speculation.

2.3.7 Is singularity inevitable?

Once an AGI is produced that is at least as intelligent as a typical human, it might appear to be something of a formality to further develop it into an ASI that goes beyond human intelligence. Having said that, reaching true AGI in the first place would be an extraordinary development, well beyond the capability of today's AI, and likely to remain firmly in the domain of speculation for many years to come.[10]

AGI can appear, to some, like a linear development from today's AI. Perhaps this is because of the functional quality of current AI tools, which give the appearance of human capabilities to perform tasks and communicate with users. However, it is important to remember how these systems work. While AI is good at identifying and categorising patterns based on how humans have identified and categorised in the past, this does not mean they 'understand' the world in a meaningful way. While they can be programmed to portray emotions, that does not mean they really feel human emotion or empathy. While they can generate interesting and useful solutions to some problems, they are not capable of creativity or imagination in the ways that are routine for humans. Crucially, while they can excel in specific tasks, they need much more information to learn than do humans, and their capabilities do not generalise well to other tasks. Even if today's AI develops at pace, one or more substantial paradigm shifts would be needed to transcend these limitations and reach the AGI that would be needed to enable 'runaway ASI'.

Nobody can predict the future – many potential AI development paths should be examined today so that we can shape AI development and prepare for the opportunities and challenges it may present. There is also an urgent need to manage and respond to the impacts that AI already has on our daily lives. These two tasks are related, but cannot be conflated entirely. For productive debate, current and speculative challenges need to be considered in the context of their methods and maturity. The following chapter aims to support this kind of debate by examining various opportunities and challenges of AI in the context of the state of the art.

[10] In Müller and Bostrom's 2014 survey of AI experts, the median estimate was a 50% chance of a high-level machine intelligence emerging by around 2045, and a 90% chance by 2075, with superintelligence anticipated some 30 years later. It has been suggested that Bostrom is over-optimistic about AI's capacity for rapid future development (if not about its impacts). Nonetheless, his 2014 prediction that an AI go player might beat the world champion within a decade was achieved within just three years. See V. C. Müller and N. Bostrom, Future progress in artificial intelligence: a survey of expert opinion, University of Oxford, 2014.

3. Why does it matter?

It is important to discuss the opportunities and challenges presented by future technologies, especially when the stakes are high, but it should be clear that context is speculative. Arguments that dismiss concerns about the future as unrealistic are common, and may be correct, but they do little to respond to the impacts of how AI is used (and abused) today. Responding appropriately to the many facets of AI will require debates that recognise the state of the art and arguments that are specific to the domains, timeframes and probabilities that they concern. To support this kind of debate, the following sections present the opportunities and challenges of AI in the context of the functionality and maturity of the technologies presented in the previous chapter. The first section focuses on the costs and benefits of today's AI, while the second explores longer-term opportunities and challenges that are contingent upon future developments that may never happen.

3.1 Current opportunities and challenges

The primary reason as to why AI matters is because of its immense potential benefits. This includes serious improvements to our health, production, mobility and decision-making, as well as indirect benefits such as efficiency gains and frivolous gadgets providing novelty or entertainment value. Even apparently inane applications can generate capital, expertise and data, which could eventually lead to more substantial developments. After all, image recognition tools perfected on cats have been redeployed to identify cancers and capacity gained in developing game-playing AI has been redeployed in healthcare. However, these technologies and their associated business practices also present legal, social, ethical and economic challenges, which will be presented in the course of this chapter. Nonetheless, since AI's benefits are the real foundation for debate about its development, the first challenge presented here is to avoid unnecessary underuse of AI.

3.1.1 Unnecessary underuse

The European Commission's 2018 communication on artificial intelligence credits AI with 'helping us to solve some of the world's biggest challenges: from treating chronic diseases or reducing fatality rates in traffic accidents to fighting climate change or anticipating cybersecurity threats',[11] while the 2020 European Commission white paper highlights its 'significant role in achieving the Sustainable Development Goals'.[12] Deploying AI to achieve such profound social value is plausible, but real-world developments are dominated by less aspirational applications, including a disproportionate focus on chatbots and efficiency tools. Even if the anticipated value of AI were limited to productivity gains, these lag far behind the market value of the firms that promise them.[13] Missing opportunities

[11] Communication on artificial intelligence for Europe, European Commission, COM(2018) 237, April 2018.

[12] White paper on artificial intelligence – A European approach to excellence and trust, COM(2020) 65, European Commission, February 2020.

[13] European Artificial Intelligence (AI) leadership, the path for an integrated vision, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2018.

to take advantage of potential benefits of AI development can be described as underuse. Such underuse can be inadvertent or deliberate, and could add up to substantial opportunity costs.[14]

Inadvertent underuse could result from failures to make the right decisions to maximise AI development and application. For example, major investment in infrastructure or research programmes that fail to yield the expected breakthroughs would, in hindsight, have been better spent elsewhere. Inadvertent underuse could also result from features of the wider cultural, economic, technical or political context that tend not to foster AI. For example, since ML is dependent on data, fragmented digital markets and a strong data protection culture could limit European development while developers operating within larger single markets with weaker consumer protection can access larger sets of data.

Deliberate underuse differs from inadvertent underuse in that it results from intentional and strategic decisions to limit AI development. This could be motivated by a desire to prioritise alignment with certain principles or values, or prompted by fears that may or may not turn out to have been legitimate. For example, decisions to limit the use of AI in medical, judicial and military domains could slow development and limit some potentially beneficial applications in the short-term. This could be justified if it avoids mistakes, promotes consumer confidence and allows time to foster a safe, effective and appropriate long-term development path. On the other hand, it could be a strategic mistake if the fears are unjustified or the opportunity costs are too great. The difficulty in avoiding underuse of AI is that it is not always clear – even in hindsight – which decisions lead to underuse and what the opportunity costs might be.

3.1.2 Four transparency challenges

Today's AI presents a range of different transparency challenges. The first and, perhaps, most salient is the lack of explainability of ML algorithms. That is, how their internal decision-making logic can be understood and described in human terms. This challenge is a function of ML methods. As explained in the previous chapter, algorithms from the symbolic AI paradigm directly reflect human expertise, and their decisions can be explained in human terms. However, an equivalent ML algorithm performs millions of calculations following their own internal logic. Even if the decisions are good quality, it is very difficult – often impossible – to explain the decision or its logic in a way that makes sense to human experts, let alone for users, policy-makers, judges and juries. This means that, even if ML algorithms make good quality decisions, their logic can be opaque. For this reason, ML algorithms are sometimes described as 'black boxes'.

While the first transparency challenge is an unintended side effect of how ML works, the second is more deliberate, as some actors exploit imbalances in access to information to serve their commercial and strategic interests. For example, ML can be used to analyse consumer data and predict individuals' 'willingness to pay' for items. Prices can be set at the upper end of the range and individual discounts sent to each shopper which – in effect – reduce the price to their estimated willingness to pay. This creates individualised pricing regimes whereby shoppers can only see the shelf price and the discount they were offered. They have no access to details of how their price was calculated, how it compares to others, or what prices are available to other shoppers. The same

[14] J. Cowls and L. Floridi, Prolegomena to a White Paper on an Ethical Framework for a Good AI Society, SSRN Electronic Journal, 2018.

approach can also be mobilised towards more strategic outcomes, as observed in political campaigns where individual voters are presented with customised messages based upon ML predictions of what would appeal to them. Citizens can only access the messages they receive, and not the full range of campaign material and promises, creating challenges of transparency and accountability. Indeed, the campaigns may even be run by third parties including commercial or foreign interests without official links to candidates. In each case, the algorithm and its owner maintain a bird's eye view while gaining commercial or strategic advantages by limiting the view of the shopper or the voter. Investigations into how algorithms make decisions – either generally or in a specific case – can be limited by measures in place to protect the intellectual property of the algorithm.

A third transparency challenge is that individuals do not always know whether they are interacting with an AI or human agent. This might include client-facing interactions such as realistic chatbot interfaces, or behind the scenes in the processing of applications for loans or jobs. Where clients are aware that they are interacting with an AI agent, they may still be restricted from understanding how and why decisions have been made, either because it is too complicated (as in the first transparency challenge), or because access is limited (as in the second transparency challenge). Users might also be unaware that they are being tracked or targeted with personalised content, whether for commercial or ideological purposes.

Artificial artificial intelligence?
A reverse formulation of this challenge has also been identified as firms pretend to use AI, hiring humans to complete tasks and claiming that they were performed by algorithms.

Fourth, and finally, there is a longer-term challenge in the lack of transparency about the full range of intended and expected outcomes of AI development. Meaningful public debate and acceptance requires transparency about the full range of expected impacts and uncertainties, both positive and negative. However, since the impacts are far from clear and there remains a high level of uncertainty, it is not easy to provide this disclosure. Hyperbolic speculation is more marketable than balanced reflection, which can easily be lost amidst the hype. Furthermore, many stakeholders succumb to the temptation of encouraging public acceptance or opposition of AI by pursuing a policy of strategic opacity about potential impacts. For example, acceptance campaigns often focus only on benefits to domains such as health while avoiding controversial but well anticipated outcomes such as the development of military and surveillance applications. These issues are further discussed in the following section.

3.1.3 Public opposition and acceptance

Public opposition is often described as a major challenge for AI. However, critical voices come from experts and stakeholders more than concerned citizens. It is possible that AI has not provoked significant visible public opposition because people broadly accept what they understand to be the costs, benefits and uncertainties. Some citizens may acquiesce to development, or feel powerless to opt out or shape it in a meaningful way. Indeed, in 2017, 61 % of Europeans were positive about robotics and AI while 30 % were negative, although 88 % agreed that these technologies need to be managed carefully.[15]

[15] Special Eurobarometer 460 (2017): Attitudes towards the impact of digitisation and automation on daily life, European Commission, 2017.

The dynamics of public opposition and acceptance could be important factors shaping AI's long-term development path. As with many technologies, public opposition to AI is often explained with reference to a 'knowledge deficit model', whereby citizens are assumed to oppose technologies because they do not understand how they work, and their concerns are interpreted as a failure to appreciate their positive impacts. Within this model, strategies for achieving acceptance include informing citizens how technologies work while highlighting the benefits they can bring and downplaying the risks. However, these approaches have been criticised as inaccurate and ineffective.[16] Inaccurate because public opposition is more often characterised by lack of meaningful engagement and control than misunderstanding, and ineffective because repeating positive messages without recognising problems can lead to a breakdown of trust and adoption of more entrenched positions. More sophisticated understandings recognise that citizens can adopt more nuanced and active roles than passive 'acceptor' or 'rejecter' of technologies. Public acceptability of AI (and other technologies) is most effectively achieved by engaging citizens early in the development process to ensure that its application is acceptable, rather than developing the technology first and then encouraging citizens to accept it as it is. Similarly, to encourage trust, it is more effective to design safe and secure systems rather than encourage citizens to have confidence in technologies that might let them down later on. How these understandings can inform action in the context of AI will be discussed in the next chapter.

3.1.4 Identifying fact and fiction

As mentioned in section 2.2.7, ML can be deployed to generate extremely realistic fake videos – as well as audio, text and images – known as 'deepfakes'. The availability of data and algorithms make it increasingly easy and cheap to produce deepfakes, bringing them within reach of individuals with relatively modest skills and resources. The deepfakes themselves are only one side of the problem, as powerful dissemination platforms – also powered by ML in some cases – can spread these materials very quickly. Together, these applications present financial risks, reputational threats and challenges to the decision-making processes of individuals, organisations and wider society.[17] The boundaries of the problem are not limited to the fake material itself. Indeed, the very existence of deepfakes introduces doubts about the authenticity of all content, including real videos. This could raise the bar for evidence – as recordings can be brushed aside as forgeries – and contribute to a broad climate of disbelief and social polarisation.

A broader problem can also be identified in differentiating between appearances and reality in the digital age. This includes reliance upon algorithms to gauge and predict performance. For example,[18] ML algorithms have been used to screen job candidates by automatically analysing videos of them speaking and using features such as speech patterns and facial movements as proxies for job suitability. Through the use of the system, these features become key metrics of predicted job performance, and can reinforce structural inequalities and biases. In a further twist, video analysis can be used to categorise candidates considered likely to have a disability, opening the possibility

[16] See P. Boucher, What if we could design better technologies through dialogue?, EPRS, European Parliament, 2019.

[17] A. Collins, Forged authenticity: Governing deepfake risks, EPFL International Risk Governance Center, 2019.

[18] Example drawn from M. Whittaker et al, Disability, bias, and AI, AI Now Institute at New York University, 2019.

of discrimination against people with disabilities or, rather, those that the algorithm defines as likely to have a disability.

3.1.5 Overcoming bias

There may be some opportunities for AI to help overcome bias. ML algorithms have been applied to look for patterns and anomalies in past decisions of judges. By highlighting the identified biases – influenced by politics and race as well as birthdays, weather and sporting results – such algorithms create opportunities to respond to them[19]. It has also been suggested that systematically applying the same algorithm to a range of cases would make it possible to ensure that the same decision-making logic is been applied consistently. This may be the case with transparent and well-designed symbolic AI systems in certain contexts, but algorithms can also reinforce and scale-up the worst excesses of human bias and prejudice.

Tay the racist chatbot

Microsoft released Tay, an AI bot that learnt to chat by analysing and engaging in conversations with humans on Twitter. Within 24 hours, Tay spoke like an angry, confused, racist misogynist. In a way, Tay is a collective failure because it acquired its unpleasantness from how humans tend to interact online. ML algorithms learn the structural biases and inequalities in our societies and find novel ways of discriminating against those that are already most affected by them.

Generally speaking, AI engineers do not deliberately produce prejudiced algorithms, but there are a few unintentional mechanisms by which they can be produced. Consider a symbolic AI algorithm for examining job applications. It might evaluate candidates by assigning scores only on the basis of their education and experience. Yet, if it fails to take account of factors such as maternity leave or to appropriately recognise education in foreign institutions in ways that human selection committees would, the algorithm might discriminate against women and foreign candidates.

Now, consider a similar AI tool within the ML paradigm. Such algorithms find their own ways of identifying which kind of candidates were selected in their training data. Where there is a history of structural biases in these selections – for example racial discrimination – the algorithm can learn these. Even where data about nationality or ethnicity is removed from the data, ML is adept at finding proxies for underlying patterns in other data such as languages, postcodes or schools that can be good predictors of ethnicity.

Several studies have shown that it is possible to de-anonymise data and make accurate predictions about individuals with reference to just a few variables. In one well-documented case, an algorithm designed to predict how likely a prisoner is to reoffend was introduced to support more objective parole decisions, but was shown to inaccurately and disproportionately discriminate against black inmates.[20] Furthermore, discrimination is also possible on the basis of inaccurate predictions, for example using algorithmic analysis of speech patterns and facial movements to 'diagnose' disabilities amongst job candidates. Across a range of domains, from justice and policing to recruitment and employee evaluation, such discrimination has combined with opacity – due to technical complexity or commercial sensitivity – to limits victims' potential for redress.

[19] D. Chen, Machine learning and the rule of law, *Computational Analysis of Law*, 2019.
[20] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown, 2016.

Another form of algorithmic bias is in the different levels of reliability for different users, with face and voice recognition algorithms consistently working best for white men. This is likely due to imbalances in the training data and can lead to unequal service provision for customers that pay for these services, as well as when these techniques are used to process data for, for example, security applications or screening job candidates.

It is important to note that AI algorithms cannot be objective because, just like people, in the course of their training they develop a way of making sense of what they have seen before, and use this 'worldview' to categorise new situations with which they are presented. The fact that algorithms have subjective worldviews is often underappreciated. They might appear objective because they apply their biases more consistently than humans. Perhaps their use of numbers to represent complex social reality gives them an air of precise facticity ('mathwashing').[21] Perhaps humans recognise their impressive power but find it difficult to comprehend their logic, so they simply yield to their apparent superiority. Whatever the reasons, recognising that AI agents are inherently subjective is a crucial prerequisite for ensuring that they are only applied to tasks for which they are well equipped. After all, if AI are discrimination machines, it is surely preferable to set them to discriminate against cancer rather than vulnerable people. Likewise, while there may be opportunities for AI to help identify biases in human decisions, it is unrealistic to expect the technology to play more than a supporting role in actions to repair deeply embedded biases in human society. These options are further examined in the following chapter.

Finally, since humans are also inevitably subjective and sometimes difficult to understand, some question why algorithms should be criticised for these same qualities. Bias and explainability do not matter much when algorithms or humans are recommending films on the basis of someone's preferences, but can be crucial when it comes to criminal proceedings or choosing job applicants. When biases become embedded in algorithms that affect these domains, they can be multiplied and rolled out at an alarming rate. We have mechanisms for dealing with humans' subjectivity and occasionally unreliable explanations. These include legal measures such as courts of law, social measures such as norms and values, and even technical measures such as lie-detector tests. One reason why algorithms are criticised as biased black boxes is because we do not always have the same safeguards in place for them.

3.1.6 Value alignment

Just as algorithms have biases and worldviews, they also have values which they continually reproduce and reinforce through their use.[22] Deviation from broadly held social values can lead to opposition and controversy, as seen with some current AI applications such as facial recognition. In response, specific values such as privacy and non-discrimination can be deliberately embedded into technologies 'by-design'. Under the symbolic AI paradigm this means programming specific instructions while, in ML, it involves controlling which data is used to train the algorithm. Value alignment can also involve limiting the use of algorithms to specific contexts and implementing robust quality control and impact assessment mechanisms.

[21] For a visual explanation, see T. Schep, mathwashing, undated.

[22] For an introduction to technology values, see P. Boucher, What if all technologies were inherently social?, EPRS, European Parliament, 2019.

Social values change, and AI values should be able to adapt along with them. Investing in technology development and deployment has lock-in effects which can also apply to values. So, while today's AI development is rightly expected to respect contemporary perspectives on autonomy and privacy, these values could take a very different shape in the decades to come. If we develop too much lock-in, there is a risk that they will gradually become misaligned as society changes. Just as it would be inexcusable to create AI with 1920s values, citizens of 2120 may feel the same about the values we deliberately embed into today's AI.

3.1.7 Informed consent: privacy, data protection and research subjects

Several concerns have been raised about informed consent for data to be stored, processed or shared for particular purposes. For personal data, the General Data Protection Regulation (GDPR) already requires informed consent for collection and processing. Finding meaningful ways of gaining informed consent can be problematic, as illustrated by the routine acceptance of terms and conditions in exchange for access to information or services. To be truly informed, consent requires that individuals fully understand what they are consenting to, so giving it is conditional upon an adequate response to the four transparency challenges described in section 3.1.2.

Informed consent is particularly important in the context of research. Here, the data subject is also a research subject. Following best practice in medical research, participants in experiments are protected by principles of informed consent. When a social media platform conducted an experiment on its users to analyse whether they could control their emotions by selecting which content to present them (result: they could), they breached the ethical principle of conducting psychological experiments without the subjects' knowledge or permission.[23] When researchers tried to identify sexual orientation from facial images alone (result: they could not), they used data scraped from data profiles, which were not intended for research purposes, not verified for accuracy and lacked consent from the subjects, they too breached the ethical principle.[24]

Verifying, identifying and classifying faces

Face recognition techniques illustrate many of the biggest challenges presented by AI. The techniques include facial verification, identification and classification. Facial verification is a one-to-one process used to check a passport picture against the face of the person presenting it, or in smartphone 'face unlock' features. Facial identification analyses images of individuals or crowds to match pictures of faces in a database. Several police forces use this to identify suspects, missing persons or other 'flagged' individuals. Facial classification is used – with or without identification – to estimate peoples' age, gender or 'emotional state'. These can be used for 'smart' billboards or lie detection, although there are serious doubts about their accuracy because people's emotional states cannot be deduced by standardised visual cues or without understanding the context. These technologies can challenge several fundamental rights. Even without their active application, knowledge that they exist can be enough to damage citizens' experience of anonymity, with chilling effects that limit their freedoms.

[23] C. Flick, Informed consent and the Facebook emotional manipulation study, *Research Ethics*, 2016, p. 12(1), 14–28.

[24] K. Quach K., The infamous AI gaydar study was repeated – and, no, code can't tell if you're straight or not just from your face, *The Register*, 2019.

AI algorithms have predicted users' sexual orientation, political views, religion and drug use from a few publicly available data points about things they 'like'[25] to inform the delivery of personalised political campaign material, and predicted whether they might have a disability to provide recommendations on job candidates.[26] The meaningful implementation of rules around informed consent is crucial to defend citizens from increasing categorisation and control by both chillingly accurate and shockingly inaccurate algorithms.

3.1.8 Security, military and dual use

While armies of robotic humanoid soldiers remain a speculative vision without basis in technical capabilities, AI does have several applications for physical combat. Before continuing, it is worth noting that military grade autonomous weapons systems are much more sophisticated and robust than many of the AI applications that citizens encounter in their daily lives. Missiles are not guided to their targets in the same way as advertisements are delivered to theirs. While critical military AI applications may incorporate elements of ML, they are better understood as highly advanced expert systems. To avoid confusion, these technologies tend to be described as autonomous systems, rather than AI.

Autonomous weapons systems such as missiles and drones can operate with different levels of autonomy. In 'human in the loop' systems, key decisions such as locking-on to targets and firing weapons are always made by a human, whereas 'human on the loop' systems operate autonomously under the supervision of a human and will continue to make key decisions until the human interrupts them. Fully autonomous 'fire and forget' systems complete their mission without any human supervision or control. While the military might not literally 'forget' about them, they would not be able to override any of their decisions. Ceding control to the system in this way reduces response time as well as the communication channels' vulnerability to detection and counterattack. On the other hand, more machine autonomy also introduces new vulnerabilities, for example to automated reciprocal conflict that escalates before human channels can intervene. Autonomous weapons systems are often differentiated by whether they are used for lethal or nonlethal missions, and whether they are deployed for defensive or offensive operations. However, the distinction is not always clear as it is often the same technology, only used in a different context with different settings. These technologies are already in use, for example in the Harpy drone, which seeks and destroys radar systems without human control or supervision.

Beyond weapons systems, AI has much wider relevance to military defence.[27] It can play an important role in both the attack and defence strategies of hacking and phishing, which have long been used for cybercrime, but can be mobilised to target key systems and infrastructures in the context of cyberwarfare. AI is also important in information warfare, for example in the use of bots to influence public discourse, threatening social harmony and democracy. At a more mundane level, the same kind of logistical AI support for supply chains can also be used to make military ecosystems more effective and efficient.

---

[25] M. Kosinski, D. Stillwell and T. Graepel, Private traits and attributes are predictable from digital records of human behavior, *Proceedings of the National Academy of Sciences*, 110 (15) 5802-5805, 2013.

[26] M. Whittaker et al, Disability, bias, and AI, AI Now Institute at New York University, 2019.

[27] See also M. Brundage et al, The malicious use of artificial intelligence: forecasting, prevention, and mitigation, 2018.

There are so many synergies between civilian and military AI that it is not possible to disentangle them entirely. As a dual-use technology, advances in civilian AI will help develop military AI, just as advances in military AI will help develop civilian AI. Taking the example of AI systems for automated piloting of drones, the same techniques designed to 'sense and avoid' in-air collisions can also be deployed for target acquisition. These synergies have existed since the earliest days of AI but, as with many technologies, the historical tendency was for military development to lead the way, followed by civilian applications. Indeed, examining the protagonists of AI's history – from Florence Nightingale to Alan Turing – many of the techniques and ideas that form the basis of contemporary AI were developed in the context of war. Today, however, the reverse of this relationship is more prevalent as civilian innovation increasingly leads the way before being adapted and applied to military applications. Revisiting the case of drones, the civilian market is highlighted as the key source of both innovation and financing for future military drone developments. Likewise, civilian AI is embraced and actively stimulated by the military as part of a deliberate long-term strategy to create opportunities for military AI. The US military research agency – DARPA – supports specific civilian AI developments with military AI as the ultimate beneficiary, and has a US$2 billion investment strategy to embed AI in weapons systems. Similarly, the US Joint Artificial Intelligence Center is tasked with accelerating the application of AI research and development across the US military. This dual use means that all AI development has implications for military AI including tools for physical, cyber and information warfare.

The specific opportunities and threats associated with military AI depend upon specific perspectives and contexts, and are subject to debate. The key advantages of autonomous weapons systems is their potential to engage in armed conflict with reduced risk of physical harm to military personnel. However, for some, lethal autonomous weapons systems (LAWS) cross red lines, for example by failing to respect human dignity.[28] As discussed in section 2.2.6, AI systems do not know what death or war is, and have no understanding of why such concepts matter. Given the synergies described above, it is not possible to entirely isolate the development of offensive, lethal and military AI from the development of defensive, nonlethal and civilian AI. As such, whichever element of militarised technology are considered unappetising – from disinformation to drones – the development and application of civilian AI can be seen as contributing to the evaluation. Recognising this, even if the dual use status of AI technology is not considered a problem in itself, it could still violate principles of responsible innovation if it is deliberately downplayed as part of a public acceptance strategy.

3.1.9 Competition: winner takes all?

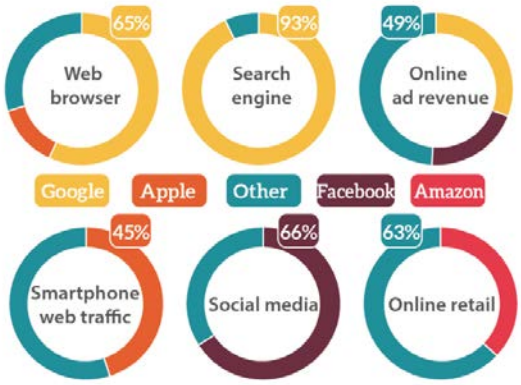
One of the transparency challenges identified in section 3.1.2 related to predictions about individuals' willingness to pay being used to fix individual prices. These practices can also present challenges to competition, but can be difficult to investigate because of imbalanced access to algorithms. It is also possible that price-setting algorithms could automatically learn to collude with each other to fix prices without the knowledge of the vendors involved. Both personalised pricing and automated collusion could be of concern to competition authorities.[29]

[28] E. Rosert and F. Sauer, Prohibiting autonomous weapons: Put human dignity first, *Global Policy*, 10: 370-375, 2019.

[29] See Pricing algorithms: Economic working paper on the use of algorithms to facilitate collusion and personalised pricing, Competition & Markets Authority, 2018.

Furthermore, the data that drives ML-enabled sectors is often collected by offering users access to services in exchange for data and exposure to advertisements. As explained in the context of spam detection in section 2.2.6, the more widely a service is used, the more data it can collect and use to improve ML services for users and advertisers alike. These services, in turn, attract more users and the cycle of data collection and service development continues. In this way, market dominance is, in itself, a driver of further market dominance.

Figure 6 – Global market share by company



Sources: W3Counter, GSStatCounter, eMarketer.

These market distortions also present a major barrier to users who consider leaving. Unlike consumers who change their internet or electricity provider, with minor cost and temporary inconvenience, those that change their social media provider lose access to the whole network, permanently. This dynamic leads to a rather extreme concentration of resources. At present, social media service providers have access to substantial information about all of their users, significant control over the information that they receive and the choices they have, and even the capability to 'nudge' their emotional states. Users, on the other hand, have limited options and few alternatives to choose from. This dynamic favours incumbents in the market which may be innovative, but can use their dominant position to outcompete or buy-out their competitors, and use their global reach to develop more tax efficient strategies.[30]

The competitive edge can also be considered on a global scale, which presents some further challenges as different global actors seek to influence how AI is developed and deployed. That is why global adoption of European values is celebrated (as seen with GDPR), while reliance upon imports is controversial (as seen with 5G infrastructure). In the 'global AI race', the EU is often positioned as struggling for the bronze medal behind the USA and China. Indeed, while Europe does maintain an important role in global AI development, particularly in terms of fundamental research, it is widely recognised that the USA and China dominate the frontline of global AI development. This is often explained in reference to their higher levels of investment, lower levels of data protection, and appetite for application and adoption. However, the 'race' metaphor is limited as it implies a definitive finish line that is the same for all participants, whereas AI is a range of technologies that actors can deploy in different ways depending on their priorities, contexts and values. Through this lens, it is most important to define the right direction and develop at an appropriate speed.

3.1.10 Distributing costs and benefits

The costs and benefits of AI are not always evenly distributed. The previous section showed how network effects can concentrate the benefits of AI in a small number of successful firms. Section 3.1.5 showed how ML algorithms learn to reproduce existing patterns of bias and inequality, so the costs of AI fall disproportionately on those that are already marginalised. Meanwhile, the benefits of

[30] See M. Szczepański, Is data the new oil? Competition issues in the digital economy, EPRS, European Parliament, January 2020.

services such as face and voice recognition consistently work better for white men than other groups.

The impact of AI on employment is regularly discussed, although much of that debate concerns the future of work and, as such, is revered for the section on speculative challenges and opportunities (see section 3.2.4). However, the platform economy is one area where AI has already had a major impact on employment, with uneven distribution of costs and benefits, whereby a new generation of digital mediators facilitate transactions between producers and consumers. Workers depend upon the platforms, but have limited access to the data or algorithms that control their activities. Furthermore, they often lack the safety nets provided by traditional employers such as regular hours, pension schemes, sick pay and family leave. Since platform workers rarely share a physical space to get to know their colleagues and are often pitted against each other through competitive rating systems, their opportunities and conditions for collective action are limited. Indeed, when they are defined as independent workers rather than employees, competition laws against price fixing could present a challenge to collective bargaining. Meanwhile, the platforms themselves benefit from network effects, with bigger players able to accumulate more data from consumers, workers and third parties and use this to consolidate their market position.

3.1.11 Dangerous overuse

Knowledge of how AI works can have a demystifying effect, revealing that some fears about AI are unfounded, but also that its true capabilities can be overestimated. While the first challenge of today's AI was its unnecessary underuse, the last is its damaging overuse.

One form of such overuse is the application of AI for tasks for which it is not well suited. This includes the use of ML for tasks where it does not excel, such as identifying causal relationships and predicting individual outcomes in complex social systems, and for those of which they are simply incapable, such as substituting for human companionship in a relationship of empathy and trust.[31] Overuse also includes the application of AI as a 'tech fix' to challenges that require a wider response. While AI may have a supporting role in responding to inequality, discrimination and inclusion, it could also distract from the need for more substantive social and economic change.

Another form of overuse is the application of AI for tasks to which it is well suited, but nonetheless lead to vulnerabilities. The regular accuracy of decision support tools can invite such levels of confidence that humans trust them automatically, perhaps more than their own judgement. As their advice is increasingly aligned with final decisions, it becomes harder to justify courses of action that contradict the AI. Experiments have shown how such 'automation biases' tempt humans to reverse their own correct decisions to align with a system that gives incorrect advice.[32] This has the effect of elevating the AI's level of autonomy beyond the supporting role for which it was designed. The implications go further than individual decisions. As humans rely upon automated tools, they miss opportunities to gain experience and refine their skills, both of which are crucial for maintaining autonomy and the capability to effectively supervise AI. This type of overuse has been cited as a major factor in several aircraft crashes. Some overuse of AI might be explained with reference to

[31] See R. Sparrow, The march of the robot dogs, *Ethics and Information Technology* 4, p305–318, 2002.

[32] K. Goddard, A. Roudsari and J. Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators, *Journal of the American Medical Informatics Association*, 19 (1) p121–127, 2012.

another bias, that is, 'bias to action', manifested as a preference for the application of available technologies despite better outcomes being achieved by simply not applying them.

3.2 Speculative opportunities and challenges

The final part of the previous chapter reviewed some speculative AI futures including self-explaining, context-sensitive, robotic, and quantum AI. The final part of this chapter presents some of the most salient opportunities and challenges that these futures present.

3.2.1 The role of utopian and dystopian scenarios

Future AI scenarios have been criticised as being too pessimistic and dystopian, although they are often unrealistically optimistic and utopian. For every premonition that AI will enslave humanity, there is a promise that it will liberate people from the need to work while taking care of our health and happiness. If the former is an unrealistic scare story, the latter is an unrealistic marketing story.

By generating hype and dominating attention, dramatic visions of future AI can have a disproportionate influence on public opinion and present a challenge in managing public expectations. However, they can also play a beneficial role. First, even if there is only a very small chance of them occurring, their potential impacts are so serious that they demand at least some reflection and preparation. Second, in extrapolating the challenges of today's AI through the occasionally dramatic lens of our hopes and fears, they provide an opportunity for broad reflection upon what we want from the technology.

3.2.2 Winter

Utopian and dystopian scenarios appear in opposition, but they share a foundation in the idea that AI will have a substantial impact on society. Their true opposite is a vision of AI fading into obscurity with little impact on society. This spectre also haunts AI in the form of an 'AI winter', a period in which interest in AI is substantially reduced, accompanied by stagnation in investment, development and application. The AI sector has experienced such winters before, in the 1970s and 1990s following periods of reduced enthusiasm in the symbolic AI paradigm that dominated the field at that time, until the resurgence of AI under the ML paradigm in the late 2000s. Some are concerned that the current 'AI summer' is merely the zenith of a hype cycle which could also descend to its nadir. This scenario is often positioned as the consequence of deploying utopian and dystopian visions which spread disappointment and fear. Others potential causes of a speculative winter have been identified in negative impacts such as unemployment and inequality, technical limits to processing capacity, and regulatory limits to the extraction and use of data.

3.2.3 Runaway artificial intelligence

One of the most commonly cited speculative challenges of AI is that it could take control of its own development, escape human control and constantly develop itself with disastrous consequences. There are several variants of this vision. In some, the AI develops its own values and aims, transcending those set by its human creators, and hides its true intentions and capabilities from humanity until they are ready to implement them successfully. Indeed, since humans are considered intelligent enough to form and pursue their own objectives, a true AGI or ASI should be able to do the same. In others, the AI sticks to the objectives set by its human creators, but with a level of capability, autonomy and determination that innocent but carelessly defined tasks such as 'create paperclips' could lead to disastrous outcomes such as enslaving humanity in paperclip factories or

transforming all earthly matter into stationery.[33] Either way, future AI's power and autonomy is presented as an existential threat to humanity. Today's AI presents no such risks, yet utopian visions of runaway AI are conspicuous by their absence. They face the same technical barriers as the dystopian variants but also require an optimistic view of the impacts of such a powerful technology, which appears beyond most people's imagination.

3.2.4 Job losses or making employment obsolete?

Job disruption is a common speculative impact of future AI.[34] In pessimistic visions, human workers are replaced by agents that do not take holidays, join unions or even draw salaries. This leads to more unequal societies as those who can perform valuable tasks or have a stake in the means of production grow wealthy while the rest face unemployment and poverty. Unlike previous waves of automation, workers lose their role in the production system and, with it, their negotiating position, leading to the emergence of an irrelevant underclass.[35] For the optimists, this job obsolescence is not a problem if the very concept of employment is also made obsolete. It has been suggested that future AI could take over almost all jobs, allowing us to build a 'Digital Athens', in which robots take the unenviable role of slaves, liberating people to occupy themselves exclusively with interpersonal, creative, leisure and sporting activities. Some might choose to work, for satisfaction or additional payment, perhaps in technology development or roles where human contact is central, such as providing social care. These two visions appear to be in opposition, but have also been combined into a single vision in which a few countries profit from AI development and provide for their citizens, while others fall behind, leading to pockets of extreme wealth and extreme poverty in different parts of the world.[36] These scenarios are deliberately provocative, compelling us to reflect upon the impacts of future AI in ways that have strong parallels with the impacts of today's AI, as presented in section 3.1.10. It seems likely that AI will affect workers from different sectors in different ways, depending upon their skills, sector, location, and ability to retrain.

3.2.5 Challenging human autonomy

Underestimating or overestimating the capability of future AI agents could be dangerous. However, assuming the development of highly capable future AI, accurate recognition of its true capabilities could also present a threat to human autonomy. For example, if we develop an AGI and we know that it makes better decisions than us, we could come to rely on it entirely for all of our decisions, effectively eroding human autonomy. In this sense, machine and human autonomy might be seen as a zero-sum game with higher quality decision support paradoxically leading to weaker human decision-making agency. The question of who follows the instructions – people or machines – is a new take on age-old questions of knowledge and power that remain unanswered with today's AI and may take on new forms with tomorrow's AI.[37]

---

[33] O. Häggström, in Chapter 5 of Remarks on artificial intelligence and rational optimism, EPRS, European Parliament, 2018.

[34] P. Boucher, What if artificial intelligence made work obsolete?, EPRS, European Parliament, March 2020.

[35] See Y. N. Harari, *21 Lessons for the 21st Century*, Jonathan Cape, 2018.

[36] See K. Lee, *AI Superpowers: China, silicon valley, and the new world order*, Houghton Mifflin Harcourt, 2019.

[37] See S. Zuboff, *The age of surveillance capitalism: The fight for a human future at the new frontier of power*, Profile, 2019.

3.2.6 Artificial emotions, consciousness and free will

If a future AGI was equivalent to a human mind, then we might expect it to have artificial equivalents to human emotions, consciousness and free will. While individuals tend to assume that other humans have these properties, today's AI clearly lacks all three. Consciousness is principally associated with awareness and experience of the self as well as the wider environment, so a future AI consciousness would need awareness of itself, what it is doing, and why. Similarly, while today's AI can sometimes implement their own decisions, they would need to step outside the paradigm of coded instructions to achieve a level of autonomy that corresponds to human free will. Emotional AI claims to reliably identify human emotions and interact with users in ways that express certain sentiments, but these functions do not correspond to genuine empathy or feeling.

Emotions, consciousness and free will have been studied and debated for millennia with fascinating yet incomplete accounts of what they are, what it means to really possess them, and how we can be sure that others have them. Nonetheless, our explicit and implicit assumptions about them underpin our lives in profound ways. Our justice system, for example, is predicated on the notion that we make choices freely and, as such, can be held responsible for our actions. Similarly, the idea that other humans are conscious beings capable of suffering (rather than being automata or hallucinations), leads us to demand the protection of other people's rights. While conscious, emotional AI with free will is perhaps the most distant of all speculative AI futures, any development in that direction – for example through brain emulation or simulated evolution – would raise several ethical issues including the possibility of AI agent's potential for suffering or entitlement to rights.

4. What can we do about it?

This chapter sets out several options that could be mobilised in response to the opportunities and challenges that were set out in the previous chapter. The first section includes policy options that shape the regulatory and economic context in which AI is developed and applied, the second highlights options for the development and application of AI, and the third focuses on social and ethical options which target the relationship between AI and society, taking account of social values, structures and processes. Each section contains seven themes with several measures each, with over 100 measures in total. They are not presented as a consolidated set of recommendations but, rather, to highlight the wide range of options for action that could be adapted and applied to specific AI developments.

4.1 Regulatory and economic options

It is often suggested that technologies such as AI should not be regulated, as it could hamper innovation. However, technology policy is often used to support innovation and, whether policy measures are taken or avoided, this should be as a deliberate and informed strategic choice to achieve a specific objective, rather than received wisdom.[38] At the European Parliament, several committees including on Culture & Education (CULT), Internal Market and Consumer Protection (IMCO), Industry, Research and Energy (ITRE), Legal Affairs (JURI) and Civil Liberties, Justice & Home Affairs (LIBE), as well as the STOA Panel, are active on AI. The Commission has set up a High-Level Expert Group, and is, at the time of publication, gathering responses to its recent white paper. Indeed, most governments, agencies and international organisations are reflecting upon their position and potential responses to AI development. While these provide some overlap, each focus on the elements that are most relevant to their mission and membership, and contribute to a flourishing global debate.

The prevailing regulatory and economic context both constrains and enables certain forms of AI development. The policy context includes primary law such as the Charter of Fundamental Rights of the European Union and secondary law such as the GDPR and Police Directive. Looking forward, the Commission work programme promises legislation for a coordinated European approach on the human and ethical implications of AI, with a proposal anticipated in late 2020. Turning to the economic context, the European case is often contrasted with the US model, as it has a more fragmented market and lower levels of venture capital. Both factors could limit the potential of small and medium-sized enterprises (SMEs) to scale-up their activities. Looking forward, the Commission has promised substantial investment in AI as well as completion of the digital single market (DSM) and new initiatives on taxation.

The following sections set out several specific options for further shaping the regulatory and economic context of AI development. However, there are also more abstract policy debates about the broad regulatory approach. This includes questions as to whether to have regulation that specifically targets AI, or to regulate it by applying and updating tech-neutral mechanisms, such as directives that apply to all products and services. Similarly, there are institutional debates about whether to set up dedicated committees and agencies for AI, or to make use those that already exist.

[38] The EU Better Regulation Guidelines require several options to be examined, including a baseline of doing nothing.

Another broad question concerns where to regulate, e.g. at Member State level, EU level, through other institutions such as the OECD and UN, or through self-regulation by actors in the AI sector. There are advantages and disadvantages to each approach, and a balance needs to be found by looking at specific challenges as well as the broader picture.

4.1.1 Create a supportive economic and policy context

The economic and regulatory environment could further support AI by reducing market fragmentation while improving consumer confidence and further supporting SMEs, start-ups and the research community. Several measures could contribute to this.

- **Complete the digital single market**. The DSM is a policy strategy to remove trade barriers for digital products and services across the EU. While the strategy has come a long way, it will remain incomplete until cross-border digital transactions are as straightforward as their domestic equivalents.[39] This is particularly important for AI applications in sectors such as transport, which need to cross borders seamlessly, and health, which require large amounts of good quality data that respects patients' dignity and privacy. Advancing the DSM in these areas may require the definition of standards for data storage, interoperability and transfer, or at least substantial increases in cooperation and coordination between Member States. The completion of the DSM could provide substantial benefits, although non-digital barriers to market defragmentation may also need to be tackled, as relatively few consumers and businesses engage in the cross-border transactions that are already available to them.[40] In the meantime, the forthcoming 'common European data spaces' will provide a system for firms to share some strategic data.[41] To succeed, the spaces need to offer value to incumbents as well as smaller firms and new operators.
- **Develop infrastructure**. Data infrastructures are crucial for ensuring the availability of high-quality data. Data trusts could help in this regard (see section 4.2.2) However, the accuracy and granularity of data is limited by the quality, width and depth of digital reach in terms of infrastructure. In this context, the development and roll-out of 5G, the internet of things (IoT), cloud computing and high-performance computing become key conditions for medium-term AI development.
- **Encourage capital investment**. The EU provides direct financial support for digital infrastructure and AI research and innovation projects and aims to stimulate a total European investment of €20 billion per year through the 2020s. The support could help to strengthen structures such as the European Innovation Council and others that promote investment in high-risk, high-return projects.
- **Support SME uptake**. Smaller firms with limited capacity may not fully appreciate the opportunities that AI presents for their business, or lack the confidence to take advantage of them. Supporting SMEs to enter the AI market – as producers, users or both – could provide a substantial boost to European AI. Digital innovation hubs already provide expertise and advice

[39] See Contribution to growth: The European Digital Single Market. Delivering economic benefits for citizens and businesses, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2019.

[40] Regularly updated figures can be found at E-commerce statistics, Eurostat.

[41] See Communication on a European strategy for data COM(2020) 66, February 2020.

to SMEs and other firms, to help them to take full advantage of digital opportunities. The AI4EU consortium was established specifically to support AI development, adoption and collaboration. These initiatives provide a good starting point, and could be complemented by information campaigns, site visits, business mentoring and other schemes targeting SMEs.

- **Articulate a development path**. For many successful AI firms in Europe, a prominent vision of future success is to be bought-out by a larger firm, often from the USA, as for example with DeepMind which was bought by Google in 2014. European AI could benefit from a reversal of this trend but, to do so, firms need the resources and confidence to develop, scale-up and mature. Alongside measures discussed elsewhere to improve access to resources – notably capital, data and skills – it could help to articulate an alternative to the buy-out vision. One approach could be to celebrate champions and to assure champions-in-waiting that support will be provided at all stages of development. Individual measures or complete programmes could be mobilised to support maturity, provided by a specialist consortium or elite innovation hub. Targeted public procurement and investment support could also play a role in helping European start-ups to mature and deliver projects with social value.
- **Adopt an ambitious vision**. Incremental developments with minor benefits for certain sectors may distract attention from more ambitious opportunities for greater disruption that could make a serious contribution to grand challenges. For example, if self-driving cars liberate drivers from the steering wheel without substantially disrupting the model, this could represent an immense missed opportunity for implementing a new generation of mobility services that offer shared door-to-door public transport while reducing the environment, health and mobility burdens associated with privately-owned single-occupant vehicles.
- **Foster mission-oriented innovation**. Achieving such ambitious visions will require not only AI development, but also substantial coordinated effort in multiple sectors and domains, and could be targeted through mission-oriented innovation. These approaches – exemplified by the Apollo programme and contemporary varieties such as the European Organisation for Nuclear Research (CERN) and the Human Brain Project – combine ambitious concrete challenges with elements of 'blue sky' exploratory research. The specific missions should be bold and widely relevant; targeted and measurable; ambitious but realistic; cross disciplinary and sectoral; and include multiple bottom-up solutions. The implementation plan can deploy a wide range of instruments – prizes, funding, public procurement and nonfinancial support – and would need to recognise the specificities of the innovation ecosystem that is relevant to the mission.[42]
- **Create innovation spaces or 'sandboxes'**. Several potentially beneficial applications – including ambitious AI mobility paradigms – lack the crucial proof of concept under real-world conditions, or require regulatory change that governments are not ready to apply. Without abandoning a precautionary approach, it may be possible to designate time- and space-limited areas for experimenting, testing and finessing specific AI applications that carry some risk but also have high potential for public good. Through a coordinated pan-European approach, it might be possible to align the spaces with mission-oriented innovation while pooling resources, sharing expertise, distributing costs and multiplying the beneficial insights across the partners.
- **Create quality labels and standards**. Labelling schemes can be used to build consumer trust and develop a brand for certain types of products. They could indicate that a product complies with EU regulations, like the CE label, or add further voluntary conditions. Existing open source

[42] M. Mazzucato, Mission-oriented research & innovation in the European Union: A problem-solving approach to fuel innovation-led growth, European Commission, 2018.

and creative commons licences already provide a form of label for software and content. Specialist labels could be designed to designate European AI, that have been independently audited, that are open for scrutiny by researchers and journalists, or that are certified to comply with specific ethics guidelines. To be effective, such labels need to be understood, trusted and used by consumers to inform their choices. Public authorities could take a lead by embedding labels in their procurement rules and conditions for access to support.

- **Engage and lead global initiatives**. While regulatory diversity can be productive and appropriate, wider coordination beyond the EU could, in many cases, be beneficial. This could be achieved by engaging and supporting wider initiatives such as the OECD work on digital taxes and the Institute of Electrical and Electronics Engineers (IEEE)'s activities on ethically aligned design. Where such initiatives are inappropriate or ineffective, the EU can lead the way, as seen when global platforms introduced GDPR compliant features not only for their EU users, but for all users worldwide. The EU could also introduce minimum standards or participation in multilateral initiatives as conditions for entering trade deals or benefitting from technology transfer programmes. However, if this approach is pushed too far or conditions are too burdensome, the problems could be exacerbated as development is pushed underground.
- **Think local**. Just as the GDPR showed how EU policy can lead the way for global action, national and even local regulations could help raise standards in wider jurisdictions. This is both because they can set an example and show what can be possible, but also because companies that want to offer universal products and services need to meet the most stringent standards in the market.
- **Amend and interpret existing law**. The Commission's regulatory fitness and performance programme evaluates the continued suitability of European legislation in light of changing contexts such as AI. They consider the relevance, effectiveness, efficiency, coherence and added value of the legislation, which can be subject to reopening and amendment procedures. [43] Policies targeting technology, such as the GDPR and Copyright Directive, are often 'tech neutral', and require guidance on how they should be interpreted in the context of specific technologies such as deepfakes and facial recognition.

4.1.2 Promote more competitive ecosystems

As discussed in section 3.1.9, the collection and use of data rewards incumbents while simultaneously increasing barriers to new entrants. Counteracting this trend through a range of pro-competition measures could help ensure a healthy ecosystem of firms providing and using AI products and services.

- **Enable consumer choice**. At present, users of social media platforms have little choice. If they want to change provider, they can close their account and lose access to the data and contacts there, and start again with a new platform which might not offer the same network. It might be possible to use the European strategy for data to build upon existing rights to machine readable data and develop open source standards for interoperable data and platforms, enabling the emergence of an ecosystem of services with more choice for users.[44]

[43] In the context of AI development, see recent evaluations of the Product Liability Directive COM(2018) 246 and Machinery Directive SWD(2018) 160, European Commission, 2018.

[44] This idea is explored in P. Boucher, What if social media were open and connected?, EPRS, European Parliament, 2018.

- **Enhance consumer power**. Measures to empower consumers to seek compensation for damages are set out in section 4.1.6 on liability. Further consumer empowerment could help to ensure that citizens' rights are respected even where the resulting damage is too minor or trivial to warrant compensation. For example, continued deliberate breaches of the GDPR that affect many people could result in compensation even in the absence of serious damage.
- **Enable data sharing**. It has been suggested that market rules have not responded to the changing reality, in particular regarding data sharing. In response, the International Monetary Fund (IMF) has advocated a clarification of the distribution of economic benefits of data, greater control for users and mandatory data sharing to weaken the position of incumbents while ensuring security and reducing fragmentation across international data markets.[45]
- **Counteract anticompetitive effects**. Because of the network effects described in section 3.1.9, incumbent firms have a major and multiplying market advantage. Measures to counteract these could include ensuring effective taxation of international firms, enabling greater consumer choice, and promoting data sharing. Authorities could develop their tools and capacity to detect and respond to market distortions such as automated collusion resulting from algorithmic price setting.

4.1.3 Improve the distribution of benefits and risks

Social inequality is a wider problem than AI but, as discussed in sections 3.1.9 and 3.1.10, the technology could concentrate profits for some while creating acute risks for others. Measures could be taken to support AI development that does not contribute to existing inequalities and, where possible, actively reverses them.

- **Develop tech taxes**. Despite the name, proposals for 'tech taxes' or 'GAFA taxes'[46] do not usually include new taxes specifically for AI systems or other technologies. Rather, they refer to adjusting the existing global tax system to ensure it is fit for purpose in the digital age. This is considered necessary because of certain features of AI and other digital products and services which make it unclear where the value is generated (e.g. in the country where the users, servers, developers, sales department or company headquarters are based). At present, the OECD is leading discussions on a global response to these issues although, if these do not yield results, EU activity could follow.
- **Redistribute value generated by AI**. Where AI and other digital products and services lead to increased inequality and concentration of wealth, other forms of targeted taxation may be considered. This could include new taxes specifically for AI-driven value generation, or more general approaches such as wealth taxes. The revenues could be distributed through existing social security systems, through innovative mechanisms such as universal basic income (UBI), or invested in measures to improve infrastructure, education, social mobility, non-discrimination and workers' conditions as part of a longer-term strategy for equality in the digital age.[47]

[45] See M. Szczepański, Is data the new oil? Competition issues in the digital economy, EPRS, European Parliament, January 2020.

[46] Referring to Google, Apple, Facebook and Amazon, considered archetypal examples of internationally dominant firms with highly efficient tax strategies.

[47] See Overcoming inequalities: investing in a more sustainable Europe, European Economic and Social Committee, 2018.

- **Protect platform workers**. AI is closely linked to many sectors of the 'platform economy' which, as discussed in section 3.1.10, has blurred the lines between employees and independent workers with platforms maintaining substantial control over the data and algorithms that define the workers' activities. The sector shows how the costs and benefits of AI can be unevenly shared. Measures ensuring a more equitable distribution of costs and benefits for platform workers could include empowering guilds and unions to represent and protect them, and to negotiate for greater social protection on their behalf. It could help to clarify how price fixing rules apply to independent workers, or how workers are defined as independent.
- **Adapt intellectual property (IP) law**. Elements of IP law, such as patents, are designed to incentivise and protect innovators by granting them time-limited monopolies as a reward for developing and publishing their creations. However, in the context of AI, this might not strike the right balance between incentivising innovation while mobilising it for social good. Adjustments could include integrating elements of open source or some form of licencing within patents to ensure that tools are available to regions and initiatives that otherwise could not afford them. Other areas of IP law such as trade secrets also present transparency challenges and could be re-examined in the context of liability and auditing of AI systems.
- **Recognise the full value of users' data**. As ML algorithms can only be developed through the use of users' data, it can be argued that the value of this data is not limited to the one-off use during the training process, but is multiplied every time the resulting algorithm is used. Without their data, the algorithm would not be the same so, in this sense, users contribute to the IP of the algorithm itself. While users are already entitled to withdraw consent for firms to use their data, they may come to expect the right to remove their data's contribution to algorithms, or to demand a share in the profits generated by their use.

4.1.4 Build resilience against a range of problematic outcomes

Developing general forms of system resilience could help ensure a more agile response to a range of potential challenges. These challenges might be similar to those associated with today's technology, such as cyberattacks and system outages, or could take a new shape, such as the speculative challenges described in section 3.2. Measures to promote resilience could also provide the auxiliary benefit of reassuring concerned citizens that precautionary measures are being taken alongside support for further development.

- **Assert the precautionary principle**. The precautionary principle is an approach to risk management which restricts specific (not general) developments that present serious risks, even if the precise details of these risks are not clearly established. It effectively replaces the burden of proof of danger with a burden of proof of the absence of danger. Its application with regards to environmental harm is already a feature of primary EU law, although a wider interpretation is commonly asserted.[48] In the context of AI, it could be applied to applications that may be considered 'dangerous until proven safe', such as the use of autonomous vehicles in crowded areas.
- **Apply impact assessment**. Impact assessments are now required before the implementation of all kinds of initiatives with potentially disruptive effects, from building works to policies. Independent assessments of specific AI applications' social, environmental, economic, health and other impacts – including any distributional effects – could be demanded before they can

[48] See the European Commission communication on the precautionary principle, COM 2000/0001.

be used in the public sector, or of applications with a high potential footprint. The impact assessment methodology could be designed to include elements relevant to values that are considered particularly important for AI, such as privacy and human dignity. Advancing this practice in the context of algorithms could help to shift the focus of development and application from efficiency gains to more direct social value.

- **Control the pace of application and development**. While slower innovation is often considered an unintended side-effect of regulation, there could be a case for deliberately limiting the application or development of some AI. Controlling the pace of application of current AI, for example, could be part of an 'innovation spaces' approach for testing and examining high risk applications. Controlling the pace of development of new AI capabilities could involve limiting research to incremental advances and ensuring there is time for public and expert evaluation of whether they cross boundaries of acceptability or create new dangers. Arguments to control the pace of development have also been articulated in the context of runaway superintelligence, as it might be easier to manage a 'slow take-off' than a rapid exponential increase in the capability.[49]
- **Collaborate internationally**. International research collaboration and policy dialogue on potentially sensitive areas such as artificial consciousness and self-improving AI would help ensure that the relationships between key actors are in place should there be need for a coordinated response to emerging challenges. Although research mechanisms already provide substantial collaboration within the EU, this could include more global state and private actors.
- **Pursue strategic technological sovereignty**. As it is considered a key enabler of the IoT – which promises to multiply the collection of data to support further AI development and deployment, particularly in transport and industrial settings – 5G is particularly relevant to AI. However, while reliance upon foreign suppliers may be the fastest and cheapest approach to roll out 5G, this could introduce new strategic and security vulnerabilities. Insisting on domestic suppliers of technology products and services would usually be described as protectionism, but such policies are enjoying something of a renaissance in the form of strategic technological sovereignty.
- **Establish (temporary) moratoriums**. Taking innovation restrictions further, some developments or applications might be banned, either permanently or until their effects are better understood. For example, temporary moratoriums on facial recognition technologies in public spaces were discussed in advance of the 2020 European Commission white paper on AI, and permanent moratoriums on autonomous lethal weapons are being hosted by the UN in the context of the Convention on Certain Conventional Weapons.[50] While establishing and enforcing bans within a specific jurisdiction may be relatively straightforward, international moratoria are fraught with diplomatic and geopolitical difficulties. Where limits are too strong, some actors might be tempted to 'go underground' in pursuit of technology development. Once risky applications are available to large numbers of people, as seen with deepfakes, such approaches may be ineffective and alternative responses will be necessary.

[49] See N. Bostrom, *Superintelligence: Paths, dangers, strategies*, Oxford University Press, 2014.

[50] See also the resolution of 12 September 2018 on autonomous weapons systems, European Parliament.

4.1.5 Enhance transparency and accountability

Many AI-based products and services are deployed via business models from the 'move fast and break things' and 'better ask forgiveness than permission' approach to legal compliance.[51] Ensuring the accountability of actors involved in AI development and application requires a certain level of transparency. Several transparency challenges associated with current AI were presented in section 3.1.2. These include the explainability of algorithms, opacity as a commercial strategy, and uncertainty about whether we are interacting with humans or AI. Here, a range of policy measures are presented, while section 4.2.2 sets out several potential technical responses to the same challenges.

- **Right to know**. Consumers could be granted the right to know the extent to which their products and services are produced and delivered by humans and AI. This includes knowing whether they are subject to algorithmic decision-making or are interacting with an AI agent, as well as whether tools that are advertised as being AI-powered are actually performed to a greater or lesser extent by remote human labour.
- **Insist upon human review**. Mechanisms could be introduced to ensure the possibility for human review and, potentially, the reversal of decisions made by algorithms.[52]
- **Demand a physical presence**. Digital firms of a certain size or activity could be obliged to maintain a physical presence in their markets, providing a point of contact for authorities and empowered consumers alike. To ensure the presence of accountable parties, platforms and mediators that enable consumers to access products and services from third parties that do not have such a presence could do so under the condition that they take on the responsibilities and liabilities of these parties.
- **Mandatory audits**. As learned through the car emission testing scandal, final product testing under user conditions is not always enough to verify compliance. Testing sometimes needs access to the inner workings of the system. Measures could be introduced for authorities to impose mandatory detailed audits of algorithms and data systems to verify compliance with legal requirements and user agreements. These could be imposed on firms or platforms suspected of noncompliance, or applied to all (or a selection of) algorithms for high risk applications. Softer measures could include requiring agreement to such audits as a condition for access to public data.
- **Public records**. Compiling and maintaining public records could respond to several challenges. The availability of public records regarding AI agents operating in specific high-risk application areas, as well as the actor responsible for them, could help respond to accountability and liability concerns. Similarly, in response to concerns about the challenges associated with personalised content, repositories of price offers, news, advertisements and political campaigns could be made accessible to all, alongside details of how their recipients were identified.
- **Protect whistleblowers**. Digital surveillance methods make it increasingly difficult to protect whistleblowers and journalistic sources. Greater protection of whistleblowers and new methods of anonymous reporting could help to reveal illegal or unethical use of data and algorithms.

[51] P. Nemitz, Constitutional democracy and technology in the age of artificial intelligence, Philosophical Transactions of the Royal Society, 2018.

[52] As argued in the resolution of 12 February 2020 on automated decision-making processes: ensuring consumer protection and free movement of goods and services, European Parliament.

- **Enhance and enforce consumer control**. Several policy measures such as the GDPR already empower users with greater control over their data. This could be further enhanced through measures to give users greater control, including transparency about how their data is used to train algorithms and how algorithms are used to process data and make decisions. Consumers of AI products should be protected from false advertising. Further measures could support consumers in exercising these rights, including mechanisms for recourse such as compensation for the misuse of data or products that do not deliver on their promises.

4.1.6 Update mechanisms for ensuring liability

Safety rules are the primary means of protecting consumers from economic or physical damage caused by faulty or dangerous products. When they fail to achieve their task, liability rules enable consumers to be compensated by the responsible party or an insurance scheme. AI products and services are no different but, for several reasons – including their complexity, opacity, autonomy and learning features – it can be difficult to prove fault and establish liability. The specific challenges of AI may also make it more important to look beyond physical damage to mental and moral damage. Having reliable liability mechanisms could help support a flourishing AI market, both by inspiring consumer confidence and ensuring better quality data. The European Commission's Expert Group on Liability and New Technologies [53] advised that the liability rules remain broadly fit for purpose, but also set out several policy options:

- **Identify operators**. For many products, the operator is clearly identifiable. However, the users of AI products might in some cases have less control over its operation than other parties such as service providers. This should be taken into account in identifying operators.
- **Make operators liable**. Operators of high-risk AI tools could be subject to strict liability – that is, held responsible for damages resulting from their use even if no specific fault or criminal intent is identified, while operators of lower risk technologies remain responsible for the proper selection, operation, monitoring and maintenance of the technology. Whether risk is considered high or low could be determined by the severity and public reach of the risk. Responsibility for damage could be maintained regardless of how much autonomy is delegated to the AI.
- **Maintain responsibility for latent defects**. Manufacturers could be held responsible for damage caused by defects even where these defects result from changes to the product that were within their control but took place after the item was put on the market.
- **Insurance and compensation schemes**. High-risk AI applications could be subject to mandatory insurance, much like private car ownership. Alongside this, compensation funds could be set up to compensate for damages that cannot be satisfied, for example because it was not possible to identify the party or technology responsible for damage. The financing of such a scheme would need to be defined, but it could provide a mechanism for satisfying many, if not most, claims.
- **Secure data**. To support the identification of faults, developers could be obliged to ensure that accessible logs of algorithmic activity are maintained securely. Failure to embed such functionality could create automatic liabilities for the producer. Under some conditions, the destruction of users' data could be regarded as damage and subject to compensation.

[53] See Expert Group on Liability and New Technologies – New Technologies Formation, Liability for artificial intelligence and other emerging digital technologies, European Commission, 2019.

- **(Do not) make AI agents liable**. The creation of direct legal liability for AI agents or robots has been suggested. This is sometimes described as 'legal personhood', but in the sense of contractual and tort liability, not rights. However, a party would still need to provide these agents with resources for compensation and, if they are insufficient, victims might pursue the same party for the remainder. As such, the approach does not resolve the issue of allocating liability, and it may also be unsatisfactory from a justice perspective. As such, the Expert Group recommended that liability should be attributed to existing persons or bodies.

4.1.7 Develop governance capacity

Public authorities and political institutions at all scales need enough expertise and capacity to respond effectively to governance challenges raised by AI. Minimally, they need sufficient understanding to make informed choices. However, some measures – such as auditing algorithms to increase transparency – could require substantial technical skills and resources. There are several options for enhancing this capacity.

- **Ombudsperson and reporting mechanisms**. An AI or digital ombudsperson could be established, empowered to audit and investigate illegal and inappropriate use of technology in the public and private sector. Similarly, a centralised mechanism for reporting and recording complaints – for example about deepfakes, problematic personalised content or unfair decision-making – could support greater understanding of the challenges as well as the formulation of solutions. New roles and procedures such as these could be created within existing structures, or by setting up new bodies.
- **Existing committees and agencies**. One option is to embed AI capacity across existing institutions and agencies. This is the current approach, as AI activities at the European Parliament span several committees including CULT, IMCO, ITRE, JURI and LIBE. Similarly, AI capacity has been introduced across several European Commission Directorate-Generals, including CONNECT, GROW and JUST. Furthermore, several EU agencies also have AI-relevant mandates, including the EU Agency for Cybersecurity (ENISA), EU Agency for Fundamental Rights (FRA), European Foundation for the Improvement of Living and Working Conditions (EUROFOUND – employment) and EU Agency for Law Enforcement Cooperation (Europol – policing), as well as the European Data Protection Supervisor (EDPS). As such, there is substantial targeted AI capacity across the EU institutions and agencies and, while there are forums for interinstitutional dialogue, there is no single body with overarching capacity or competence.
- **Dedicated institutions**. At the time of publication, the European Parliament is setting up a special committee on artificial intelligence in a digital age, with a 12 month term of office.[54] Two European Parliament resolutions have called for the creation of a dedicated EU agency to ensure the capacity to monitor and respond to AI development.[55] The European Commission has stated

[54] See decision of 18 June 2020 on setting up a special committee on artificial intelligence in a digital age, and defining its responsibilities, numerical strength and term of office, European Parliament.

[55] 'Civil law rules on robotics' called for an agency for robotics and AI and 'A comprehensive European industrial policy on artificial intelligence and robotics' called for an agency for algorithmic decision-making. See resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics and resolution of 12 February 2019 on a comprehensive European industrial policy on artificial intelligence and robotics, European Parliament.

that it does not consider a new agency to be necessary, citing existing instruments, dialogues, strategies and resources. However, if options such as algorithmic auditing, reporting systems and public records are adopted, or specific expertise is required to define whether a given application is high risk (a status which could change the applicable liability rules), an agency charged with these tasks could become increasingly attractive. At an international level, governance capacity could be supported through new dedicated institutions, for instance, an 'IPCC for AI'.[56] While such dedicated institutions could allow for a more comprehensive approach to AI policy, they would not replace the targeted competences of others (in industrial strategy, fundamental rights, etc.), nor the need for substantial dialogue and broad coordination with those bodies.

- **Third parties**. Governance capacity could also be supported by setting-up and supporting third parties to monitor developments, shape debate, and advise on how to respond to emerging opportunities and challenges. This could be achieved through observatories or, less formally, through a range of individual projects.
- **Other public sector capacity**. While most public sector bodies might benefit from developing AI capacity to take advantage of new opportunities, some – such as justice and media – need some degree of AI capacity in order to continue fulfilling their basic mission in the digital era by scrutinising and actively responding to AI developments.

4.2 Technology development and application options

This section presents options that rely principally on the application of technology in response to the opportunities and challenges presented by AI. These measures are only partially distanced from the policy options described in the previous section, first because enacting them may require soft or hard regulation, and second because technology can be 'politics by other means'.

4.2.1 Technology values

Technologies script, enable and constrain different behaviours and relationships and, in doing so, they promote conformity or deviation from different values. They can shape personal values such as how much information we share online, as well as political values such as how far we centralise our services and authorities. The values themselves can be embedded unintentionally or deliberately, and can be made durable through 'lock-in' effects. Several measures can be taken to shape technology values. Ideally, they would be subject to broad debate and implemented with a degree of consensus, as discussed in section 4.3.5.

- **Values-by-design**. 'By-design' strategies aim to embed desirable qualities in technology from the earliest stages of their development. This is considered more elegant and robust than adapting technologies that are already in use to comply with legal standards or social expectations that they fail to meet. As such, privacy-by-design refers to orienting the whole technology development process around ensuring that privacy will be respected throughout the technology's lifecycle. This could mean, for example, limiting the potential for deliberate or accidental sharing of sensitive data, or guaranteeing the timely deletion of data that is no longer needed.

[56] See N. Miailhe, AI & Global Governance: Why We Need an Intergovernmental Panel for Artificial Intelligence, United Nations University Centre for Policy Research, 2018.

- **Beware value lock-in**. As technologies become embedded within social structures and practices, they script social activities and become, along with the values they represent, reference points for normality. Once established, they are difficult to change, as observed for example with individual car ownership. It is important to pay attention to what kind of activities are being normalised by AI and their wider effects. This reflection is beginning to happen with regards to the free provision of online content and services in exchange for data and targeted advertisements. Quicker reflexes have been observed this year on developments such as facial recognition and could be sharpened through sustained high-quality debate amongst policy-makers, civil society organisations, industry and experts.
- **Interpret and embed values for AI**. There are many proposals for identifying and interpreting values for AI. These are explored in section 4.3.2 on ethics frameworks. However, since few disagree with broad principles such as justice and human dignity, agreeing on general values can be straightforward. However, defining what they mean for AI development in concrete terms requires their translation into specific priorities and requirements that either restricts some AI technologies or applications, or shapes how they are implemented from a technical perspective. Some major initiatives support this task, including the European Commission's High-Level Expert Group on AI, which is testing and piloting 'assessment lists' designed to help developers to implement guidelines.[57] The IEEE is overseeing an initiative to deploy standards, certification programmes and global consensus-building to ensure AI developers are educated, trained, and empowered to build ethical AI, and has published a substantial compendium on how to do so in a wide range of contexts, from autonomy to intimacy.[58]
- **Allow for changing values**. Values change over time. What is unacceptable now might be embraced in the future, and values that we work hard to embed in today's AI might be unpalatable to tomorrow's citizens. While it remains sensible to embed today's values in today's AI, this should not hinder future generations from reinventing concepts such as privacy or autonomy for themselves. As such, alongside values-by-design, 'flexibility-by-design' could help ensure that future generations are not tethered to today's values. This could be achieved through measures to reduce financial and environmental lock-in, such as supporting adaptable open source software, ensuring the presence of healthy ecosystems of technology companies and applications, and using more reusable and recyclable components. Social reflexes on these issues could be developed through reflection on the role and impact of technology in society.

4.2.2 Accessible data and algorithms

As set out in section 3.1.2, today's AI presents a range of transparency challenges including the explainability of algorithms, opacity as a commercial strategy and uncertainty about whether we interact with humans or AI. Policy measures responding to these challenges were set out in section 4.1.5. Here, technology measures are presented that could make data and algorithms more accessible to a range of actors, including authorities, users, researchers and developers.

[57] High-Level Expert Group on Artificial Intelligence, Trustworthy AI assessment list, European Commission, 2019.

[58] See the Global Initiative on Ethics of Autonomous and Intelligent Systems, Institute of Electrical and Electronics Engineers, undated.

- **Comment the code**. Software engineers are traditionally trained to add comments to their code to explain what each part of the software is doing and how it does it. While some argue that good code explains itself, good comments (or, where appropriate, separate manuals) will improve accessibility not only for other developers, but also for auditors and others that may need to understand the operation of the algorithm in future.
- **Make use of explainability mechanisms**. While self-explaining AI remains a speculative technology, there are some tools for examining how today's ML algorithms make decisions.[59] These include assessments of which data inputs are particularly important in shaping algorithms, which features are particularly important in shaping their decisions, and what kind of changes of input would be required to provoke a change in the algorithm's decision. These tools are not particularly user-friendly and are most often used for 'debugging' purposes, although they could be used to help developers understand how their algorithms work, for expert auditors to check compliance, or for researchers to explore the potential impacts of their application.
- **Support open source and creative commons**. Initiatives for accessible data and algorithms such as open source and creative commons licences can be supported through public procurement and research funding. They might also benefit from greater protection from illegitimate use, for example when algorithms or content are taken and used for commercial purposes without respecting the conditions of their licence.
- **Enable portability across platforms**. Creating a competitive market for services such as social media could give individuals more control over how their data is used and foster more responsive and accountable practices in the sector. The GDPR gives citizens the right to obtain readable, portable copies of data about them that is held by platforms. A further measure could be to support open standards and open source development towards the emergence of a new generation of interoperable platforms, enabling users to change platforms seamlessly without losing access to their contacts or data.[60]
- **Open the APIs**. Application programming interfaces (APIs) are the access points that applications use to engage with larger platforms and systems. More open APIs enable third parties to access data (e.g. content stored on a social media platform) and produce software (e.g. applications that work with a system), while more closed APIs maintain greater control for the owner of the platform or system. Most APIs have strict access limits and require pre-approval, even for projects by established researchers and authorities conducting investigations. Some researchers and public authorities have resorted to 'scraping' content by browsing platforms with regular user accounts, although this is against the terms of service and data may be unlawful, unethical and low quality. Measures to enable greater access for researchers and public authorities may be valuable in ensuring that AI development is trustworthy and legal. This could be achieved through more open APIs – particularly for researchers, journalists and public authorities – or through more radical reconfigurations of how data is controlled.
- **Set up data trusts**. To improve trust and respond to uneven distribution of the benefits and risks of sharing data, data trusts could be set up. These are legal and technical structures that enable large amounts of quality data to be used to develop AI while defending the rights, preferences

[59] See U. Bhatt et al, Explainable machine learning in deployment, Proceedings of the 2020 Conference on Fairness, Accountability & Transparency, January 2020.

[60] See P. Boucher, What if social media were open and connected?, EPRS, European Parliament, 2018.

and interests of those that provide it.[61] Pilot studies are underway with a range of case studies and, depending on the results, could be taken further through public-private partnerships. These could start with low-risk, high-benefit data provided by different authorities, such as traffic and pollution indices, before moving on to more personal data provided by users about their health or habits. Access to trusts can be made conditional upon compliance with transparency and accountability principles. Where their data is sufficiently valuable, they could provide substantial leverage to promote responsible digital practices.

4.2.3 Quality data and algorithms

In response to the challenges of bias in ML algorithms, set out in section 3.1.5, various means of removing biases from training data have been suggested. These measures are part of a wider impulse to improve the quality of data and algorithms. Some applications need near perfect quality. For example, a mass screening tool to identify suspects in busy public places that is 99.99 % accurate would still lead to dozens of false positives per day. Good quality algorithms need good quality training data. However, comprehensive quality data about society will also reflect that society's biases and inequalities. As such, until these biases and inequalities are reduced in society, technology options are limited to recognising and highlighting their presence in data and algorithms, and minimising their effects through safeguards and judicious application.

- **Avoid 'garbage' data**. The old computer science maxim 'garbage in, garbage out' grows truer by the year. It refers to the quality of the inputs such as training data and the quality of the output such as algorithms and their results. Huge amounts of data can be 'scraped' from social media and various other sources but – while cheap and readily available – they are not always representative, accurate or fit for purpose. Social media content and location data are increasingly used as proxies for social realities, for example to 'map' social perspectives or 'track' physical contact between people. However, this data is often unsuitable for these purposes. Social media accounts are not representative of society (indeed, many are simply bots) and their power as a gauge of public perspectives is regularly overestimated by news organisations and political actors. Similarly, location data can be unreliable and ignores key features such as walls and floors. The use of 'data in the wild' might also breach rules about copyright, informed consent or the processing and storage of banned material. Standards could play a role in improving the quality of data. The FAIR format (findable, accessible, interoperable and reusable), for example, could be applied to new and legacy data to support quality control.
- **Remove data labels**. Google's API for image recognition no longer includes gendered categories. So, for example, images of engineers do not have gendered labels such as 'man'. It has been suggested that removing labels in this way could help reduce algorithmic bias, for example against women in shortlisting candidates for engineering roles. However, the power of ML is its ability to find its own novel ways of predicting trends in data. If gender was a discriminating factor in previous human decisions, the algorithm can learn to predict this feature before using it to discriminate. The risk is that, by removing the labels, structural biases in data and algorithms are not removed, but simply hidden. Where bias in data and algorithms is identified and cannot be genuinely removed, it may be better to highlight its presence and limit how it can be used.

[61] An accessible introduction and exploration is provided by Element AI and Nesta, 2019.

- **Assess data quality**. Tools can be applied to identify quality issues such as incorrect data labels, inappropriate biases, illegal material (including information gathered without consent) or 'fake news'. Depending on the system, material might be removed automatically or flagged for human review. The difficulty with such tools is that they are also biased, in this case against data that does not conform to their definition of quality. Definitions of quality reflect perspectives which are not always universally accepted. This is particularly true of such material as campaign messages, news and misinformation. As such, these tools need to be developed cooperatively and continually tested to avoid manipulation or overuse.
- **Recognise the limitations**. Data that reflects human decisions in domains that feature structural biases cannot be complete, accurate and unbiased at the same time. In these cases, it is important to be aware of these biases and ensure that algorithms are not used in domains and functions for which they are not well suited.

4.2.4 Apply with care

There are technical reasons why AI should not be used to perform certain tasks. While AI can be good at pattern matching and identifying broad statistical correlations, it is not equipped to perform other tasks such as predicting individual social outcomes. Indeed, some of the most damaging examples of the misuse of algorithms come from the use of algorithms for tasks for which they are not well suited, such as predicting whether an individual will reoffend or perform well at work.[62] On a wider scale, embedding AI-enabled systems in our infrastructures could introduce new vulnerabilities. At present, citizens are most directly exposed to functioning AI in content distribution, usually designed to sell products and ideas. The case for promoting AI would be stronger if its development was mobilised to provide profound and tangible social good rather than minor efficiency gains, particularly when the costs and benefits are unevenly distributed.

- **Limit some technologies or application domains**. Domains such as justice, policing and employment have been highlighted as inappropriate for the use of AI. However, not all AI applications within these domains are risky. Within justice, for example, there are many uncontroversial applications, such as supporting case law analysis or access to law. The European Commission for the Efficiency of Justice[63] differentiates between uses that should be encouraged, that require considerable methodological precautions, which should be subject to study, and should only be considered with the most extreme reservations. Similarly, controversial AI techniques such as facial recognition have been flagged as fundamentally unacceptable in contexts such as mass identification in public places, but acceptable in others such as identity verification to unlock phones.
- **Adopt a risk-based approach**. There are many ways of defining which applications are high risk and what measures would apply in these cases. The European Commission is currently examining the definition of high-risk applications with reference to the specific application as well as the sector in which it is deployed, paying particular attention to how these two factors may combine to present serious risks to citizens. Some applications (such as the use of biometric

[62] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown, 2016.

[63] See the European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment, European Commission for the Efficiency of Justice (CEPEJ), 2019.

data for remote identification) or sectors (such as recruitment processes) might always be considered high risk. Uses of AI that are considered high-risk could be subject to greater burdens such as impact assessments and strict liability rules, while those deemed to have unacceptably high-risk might be subject to temporary or permanent moratoria.[64]

- **Assess systems vulnerabilities**. Embedding AI in our basic infrastructure could create new forms of systems vulnerability. Europe is currently setting up 5G networks to support an immense network of connected devices in an 'internet of things' (IoT), which will enable new dimensions of industrial and social data production. As we become reliant on these networks and the new data-driven services they enable, we may become more vulnerable to disruptions caused by energy shortages, cyberattacks and other unexpected failures and side-effects. These vulnerabilities could be counteracted by measures to increase resilience by avoiding lock-in and maintaining strategic autonomy in resources and expertise.
- **Prioritise applications with real social value**. While there are frequent references to the profound benefits of AI, most citizens' practical experience of AI is limited to relatively frivolous benefits while, for firms, deployment is disproportionately focused on chatbots and efficiency tools. Even there, implementation lags behind the promises.[65] Furthermore, the benefits disproportionately accrue to providers and those in a privileged position while costs fall upon those who are already marginalised. Continuation of this trend could lead to disillusionment. To foster broad support for AI development, it may help to prioritise applications that respond to grand challenges and provide genuine, clear and direct benefits to individuals' health, environment, work and personal life. Some of the measures outlined in section 4.1.1 could be valuable here, including adopting an ambitious vision, articulating a development path and fostering mission-oriented innovation to achieve it. It could also help to provide more open access to information, share control of processes, and ensure a more even distribution of benefits and risks.

4.2.5 Use available 'tech fixes'

A 'tech fix' is often an often derogatory term that refers to technological solutions to problems, usually those created by the application of other, prior, technologies. However, there are some cases when AI could be deployed to help respond to challenges presented by AI.

- **Automated flagging**. AI tools can be used to identify instances of abuse, such as breaches of the law, deepfakes, spreading of mis- and disinformation and cybercrime. Identification could trigger responses on the supply side, such as automatic removal of content or notification of human supervisors, or on the user side by providing information about the problem and how they can respond to it. Such tools can be used to identify and alert users about unfair clauses, for example in the terms and conditions of digital products and services. As the volume of data traffic increases, such tools are increasingly presented as the only feasible means of effective monitoring. However, at present, they are primarily used by employers and suppliers rather than users and the standards that they enforce might not always be transparent or universally accepted.

[64] See the White paper on artificial intelligence – A European approach to excellence and trust, COM(2020) 65, European Commission, February 2020.

[65] European Artificial Intelligence (AI) leadership, the path for an integrated vision, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2018.

- **Kill switches**. Algorithms could be designed with embedded mechanisms for humans to immediately halt automated activities at any moment. Such kill switches or 'big red buttons' are standard practice for robotic systems that present risks of physical damage. Their adaptation to software would need to be accessible for legitimate use yet secure from malicious use, and include features such as secure storage of logs detailing the system state and processes ahead of the termination. One problem with such tools is that they are designed for use once serious problems become visible, which may be too late. As such, kill switches need to be seen as a last resort mechanism to complement continual monitoring and preventative measures.
- **Recalibrate personalisation**. Some forms of personalisation such as political advertising and individual pricing have potentially damaging side effects, as set out in section 3.1.2. Several technical responses are available, including switching them off, making them optional, allowing users to choose profiles for themselves, and maintaining fully transparent and accessible catalogues of price offers, campaign material and other personalised content along with their target audiences.
- **Red teams and white hats**. 'White hats' are hackers that seek to identify vulnerabilities so they can be fixed, and they can form part of 'red teams' that are deployed to attack or criticise systems with the aim of improving them. Such teams can include AI tools that are already in use for malicious purposes and could be mobilised for constructive criticism of AI systems and applications.

4.2.6 Develop speculative 'tech fixes'

Taking the concept of tech fixes further, it may be possible to develop new capabilities, beyond the reach of today's AI, in response to both the current and speculative challenges of AI.

- **Self-explaining AI**. As introduced in section 2.3.1, work is underway to try to develop AI that is capable of explaining its decisions in an accurate and understandable way. At present, the detailed configuration of an ANN provides the only complete explanation for its output. However, these explanations are too complex, abstract and time-consuming to satisfy humans. Dependable self-explaining AI could reveal its own errors, biases and limitations, responding to transparency challenges set out in section 3.1.2 and making them more suitable for critical decision support roles. This could support the verification and improvement of systems and also enable examination of compliance with legal requirements and social values. While such a development appears modest in comparison with 'blockbuster visions' of future AI, it remains outside the current paradigm of AI development and may remain elusive in the medium- to long-term.
- **Social AI**. It could be possible to develop a new paradigm of AI that is oriented around AI's role in supporting humans and being well embedded in social systems.[66] Its algorithms and decisions would be contestable from the ground up, and the results would not present final decisions, but instead a range of options along with their associated impacts (with probabilities) and the underlying logic of the results.
- **Self-identifying AI**. A speculative tech fix to the transparency challenge of knowing when we are interacting with AI agents, as well as other problems associated with unidentified AI outputs such as deepfakes, could be to create a kind of watermark that permeates AI systems' outputs to

[66] See H. Fry, *Hello world: How to be human in the age of the machine*, Black Swan, 2019.

identify their origin to users and other systems. The watermark would need to suit the format of the output, e.g. sound, text images, videos and algorithms composed by AI.

- **Walled-off AI**. In response to the speculative challenge of runaway AI, it has been suggested that AI could be 'walled-off',[67] limiting its interactions with the world. This could be achieved by completely separating it from other infrastructures such as the internet, limiting it to one-way information flows, or by restricting it from producing executable commands. In response to concerns that a really advanced ASI could outsmart any mere human intelligence, it has been suggested that such a system could be designed as an oracle that can only respond to specific questions.

4.2.7 Constructive reflection and critique

Social problems such as structural bias and inequality have been around for much longer than AI, and it is unfair to expect technology to resolve them. Before moving on to the final section, which sets out some possible societal and ethics measures, this final set of technical options focus upon critical reflection about the capabilities, development paths and impacts of AI.

- **Productive critique**. For traditional scientific communities, critical debate involves engaging specialist knowledge communities in almost masochistically exploring weaknesses in understandings, methods and products, and improving them in response. It has been argued that critical discussions about AI are designed more for wider audiences and lack the level of reflexivity and epistemic modesty needed to make them productive.[68] Much of the AI development community has a more commercial than academic culture but might reap productivity benefits by borrowing the norms of academic debate.
- **Epistemic modesty**. Basking in the limited achievements of AI – and exaggerating its capacity to resolve major problems such as the spread of Covid-19 – can contribute to hype cycles and promote damaging overuse, potentially misdirecting resources from more productive responses. To help reverse these trends, participants on all sides of the AI debate could practice greater modesty regarding its capabilities (e.g. claims of creativity or hostility) and impacts (e.g. on our health or employment).

4.3 Societal and ethics options

The options presented in this section are principally oriented around social processes or values, even though they might be implemented through policy or technology measures. Many are less narrowly targeted than those described in the previous sections, responding to challenges that are relevant to, but broader than, AI.

4.3.1 Skills, education and employment

Several measures could target skills and employment in response to both AI's opportunities and challenges. This includes ensuring that young people have the appropriate skills to navigate their personal and professional lives, that society has the skills and capabilities needed to exploit the

[67] See S. Armstrong, A. Sandberg and N. Bostrom, Thinking Inside the Box: Controlling and Using an Oracle AI, 2012.

[68] See H. Collins, *Artifictional intelligence: Against humanity's surrender to computers*, Polity Press, 2018.

opportunities of AI, and that AI does not contribute to but rather reverses inequalities including digital divides, social exclusion and the uneven distribution of costs and benefits.

- **Computer science and programming on the curriculum**. Learning AI, computer science and programming skills could help many students to navigate their careers and also to manage their increasingly digital lives. These disciplines could be introduced at an earlier age and for a greater portion of time, and could also be combined with other disciplines. For example, programming elements could be included in sciences and humanities studies. AI training could also be embedded in the university curriculum for the next generation of lawyers and medical doctors, for whom AI is likely to play a key supporting role in their future careers.
- **Transferable skills**. As job markets are anticipated to change more regularly and radically in the future, the next generation of employees may benefit from learning more transferable skills, in particular adaption and 'learning to learn'. This could be achieved through a greater focus on skill acquisition and problem solving in school curricula.
- **Continued learning for employees**. Substantial retraining for mid-career workers would help employers and employees alike to manage transitions in the nature of work and the skills required to flourish. However, once people start their careers, further education is usually limited to either very short courses offered by employers or longer programmes targeting unemployed people. The concept and delivery of continued learning could be renewed to support more proactive retraining that anticipates changing needs during employment. This could include creating new ways of delivering, certifying and financing mid-career retraining that is delivered 'on-the-job' with support from universities and professional institutes.
- **New career roles**. It is easier to see how AI can displace current jobs than it is to imagine those that it may create. However, when new roles do emerge, support could be offered to help them develop them into established career paths. Take, for example, a new type of job role as a digital advisor that helps individuals to manage their privacy settings, to hold service providers to account, and to understand new risks, liabilities and opportunities. To flourish, this career path would require a particular blend of legal, technical and communication skills, as well as broad recognition and trust. This could be supported by engaging with the vocational and professional training sector to develop bespoke certification and skills development programmes, establishing new professional bodies, through public procurement, and by offering new public services to citizens.
- **Advanced research**. European universities are considered to be strong in AI, but their position could be enhanced by support for networks and major collaborative projects such as those discussed in the context of mission-oriented research and innovation spaces. Universities find it difficult to compete with the salaries and research projects that are offered in the private sector. New models could be established to enable researchers to maintain strong links with universities while still pursuing private sector careers. Skilled researchers could be encouraged to move to Europe through mobility schemes including visa and mobility support, particularly for those coming to work in a university. Collaborative links could also be strengthened through public-private partnerships that engage universities and private sector actors (SMEs as well as larger firms) to deliver products and services with high social value.
- **Professional codes and standards**. Given the impact of their work, AI professionals may one day be considered to have similar responsibilities and duties of care as medical doctors. While publicly funded research may be subject to clearance from independent ethics boards, cutting edge AI development is more often found in the private sector, where such boards are not in place. Action could be taken to foster a culture of responsibility amongst developers, perhaps

via membership of professional bodies or through the development of a 'Hippocratic Oath' for AI developers, whereby developers endeavour to uphold certain principles and professional standards.

4.3.2 Apply ethics frameworks

In recent years, a proliferation of ethics principles and guidelines for AI have been published.[69] While the proposals vary to some extent, European initiatives tend to follow the lead of the EU Charter for Fundamental Rights, which contains several AI-relevant rights and freedoms[70] while drawing further inspiration from wider sources such as medical ethics[71] and literature.[72] The Commission's High-Level Expert Group on AI asserted that trustworthy AI should be lawful, ethical and robust, setting out seven requirements to achieve this.[73] The level of convergence amongst proposals has been described as cause for optimism, as it illustrates broad agreement about how AI should develop. However, the consensus is limited by the level of flexibility in interpreting and implementing such principles. The following societal and ethics measures to support ethical AI complement those technical measures set out in section 4.2.1.

- **Develop rights for the digital age**. To close digital divides, it could help to create de facto or even formal rights to access digital products and services. This could mean programmes to provide internet access or basic computing services. On the other hand, it could be beneficial to develop rights to withdraw from the digital sphere. This could mean guaranteeing the provision of face-to-face provision of public services, or limiting the use of face categorisation in public spaces. The two rights could complement each other quite well to ensure that citizens can flourish in the digital age. There have also been proposals for 'new rights' to meaningful human contact or for protection from profiling, measuring, analysing, coaching and nudging.[74]
- **Shift from general to specific**. Building consensus around the notion that AI should respect broad principles such as human dignity has proven relatively easy. However, to make them

[69] For a recent review, see The ethics of artificial intelligence: Issues and initiatives, EPRS, European Parliament, March 2020.

[70] On one hand, automated decision-making could threaten rights to human dignity, non-discrimination and good administration; the collection and processing of personal data could challenge rights to respect for private life and data protection; and new models for news production and consumption could restrict freedoms to hold opinions and to receive and impart information. On the other hand, citizens' right to develop AI applications and bring them to market without disproportionate restrictions are protected by freedoms to conduct business and undertake scientific research.

[71] The classic principles are autonomy, justice, beneficence, and non-maleficence, although human dignity, privacy, epistemic modesty and others also feature in the contemporary Hippocratic Oath.

[72] For example, the three laws of robotics devised in works of science fiction by Isaac Asimov state that AI agents should value their own existence as long as they value human wellbeing and orders more. Others argue that it is unsafe for AI to value its future existence at all. However, since today's AI is not capable of such reflection, the ideas are difficult to evaluate and are somewhat moot.

[73] Human agency and oversight, technical robustness and safety, privacy and data governance, transparency, societal and environmental well-being, accountability and diversity, non-discrimination and fairness. See High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, European Commission, 2019.

[74] Statement on artificial intelligence, robotics and autonomous systems, European Group on Ethics in Science and New Technologies (EGE), 2018.

operational, they need to be translated into specific measures, which may reveal how certain processes and applications diverge from principles and could lead to the collapse of consensus. To manage the shift from general to specific guidelines, ethicists could work with developers to explore the possibilities and examine their effects. As discussed in section 4.2.1., there are some good initiatives in this direction although their impact is limited by their voluntary nature.

- **Shift from voluntary to binding**. Aside from those elements already established in law, adherence to ethics frameworks remains voluntary. When firms define their own codes, it is difficult to establish whether they make a substantial practical difference, and there are no mechanisms for enforcing their own compliance, or for ensuring that adopting principles does not create a competitive disadvantage. If sufficient industry-wide principles cannot be achieved, binding legal measures can be developed. However, these would require even more specific interpretation of principles as well as penalties for noncompliance, raising the stakes and risking further degradation of consensus. It could help to reorient discussions about AI ethics to AI rights, as the latter already have legal force, albeit in general terms, and are often closely linked to ethical principles.
- **Establish digital ethics committees**. Following the example of bioethics committees at institutional, national and international level, AI or digital ethics committees could be established to advise governments and provide an interface with international organisations, research councils, industry bodies and other institutions.
- **Integrate ethicists meaningfully**. To ensure the conformity of products with ethical principles, it has been suggested that AI ethicists could be embedded into firms and development teams. However, the effectiveness of such an approach depends upon the roles to which they are assigned and the priorities of the activity. Ethicists can be employed for 'ethics washing' and managing reputational risks, and their influence may be limited to 'low-hanging fruit' and 'win-wins'. To succeed, ethicists would need to be deeply embedded in development teams and have enough technical expertise and management support to make a difference.
- **Consider moratoriums carefully**. Moratoriums have been suggested in response to ethical issues presented by AI applications that are already in use (such as facial recognition), that are technically feasible but not currently in use (such as fully autonomous lethal weapons), and that are purely speculative (such as artificial consciousness). Temporary restrictions could allow time to examine the impacts and options, while permanent bans aim to outlaw applications or stop them from being developed in the first place. Whatever the approach, moratoria rely upon widespread consensus and trust. Without them, development may simply be pushed underground or relocated to jurisdictions that are unable or unwilling to enforce the ban.

4.3.3 Greater workplace diversity

The AI sector has been accused of having a gender, race and disability diversity crisis.[75] A more diverse and representative workforce could help it to produce less biased algorithms while contributing to the alleviation of wider structural inequalities and biases.

- **Promote more diverse workplaces**. Several measures have been proposed[76] to improve workplace diversity. These include greater transparency about pay, recruitment, promotion, and

[75] M. Whittaker et al, Disability, bias, and AI, AI Now Institute, 2019 and S. M. West et al, Discriminating systems: Gender, race and power in AI, AI Now Institute, 2019.

[76] S. M. West et al, Discriminating systems: Gender, race and power in AI, AI Now Institute, 2019.

cases of harassment and discrimination, greater equality in pay, recruitment and promotion, wider recruitment drives, improved pathways for workers that are not full-time employees, and linking executive incentives to hiring and retaining under-represented groups.

- **Recognise and reduce barriers**. When traineeships and entry-level positions are not sufficiently remunerated they present a barrier to those that do not have enough resources to participate. AI firms could explore the barriers faced by under-represented (prospective) employees and take action to reduce them.
- **No additional burdens**. While greater workplace diversity might help respond to the AI sector's bias and inequality challenges, the task of finding solutions to problems should not fall disproportionately on those that are already marginalised the most by them. So, for example, a female software engineer doing the same work as her male colleagues should not, in addition, be expected to find new ways of resolving gender bias in algorithms or hiring practices. Diversity helps, but these tasks will likely require full-time attention by specialists with qualifications that go beyond membership of the affected group.
- **Counteract sectoral stereotypes**. Stereotypes of the AI sector's demographics are not entirely unfounded, so it would be misleading to 'diversity wash' the sector and pretend it is more representative than it really is. However, some measures could be deployed to counteract this image in order to encourage wide participation. This could include, for example, avoiding all-male panel discussions and ensuring diversity amongst its representatives at meetings and conferences.
- **Ensure genuine inclusion**. More diverse workplaces will enable firms to draw upon a wider range of experiences and insights. However, to benefit from this resource, diversity initiatives need to go beyond the numbers to ensure that those currently under-represented in the sector have the skills, experience, authority, resources and confidence to influence development.
- **Reflect and act upon wider structural inequalities**. Measures to increase diversity in the sector may have limited effects if they are not accompanied by wider reflection and action on the reasons as to why it lacks diversity in the first place. Overcoming the diversity crisis will require changes in attitudes and practices across the sector and, indeed, society more broadly.

4.3.4 Improve inclusivity and equality

Broad social inequality and exclusion are the fundamental cause of some of the AI sector's key challenges. They drive its diversity crisis and – since quality data about society will inevitably reflect structural inequality and exclusion – its algorithmic biases. Effective remedies to these underlying issues might help the AI sector to respond to its own bias and inclusion challenges. However, while the sector can play its role, it cannot be expected to resolve these problems alone. Wider measures to respond to equality and inclusion challenges are subject to broad political and social debate that fall out of the present scope. Nonetheless, the following passages highlight some measures of particular relevance for AI.

- **Reduce digital divides**. Not all groups have equal access to digital services and their benefits. This includes infrastructure (such as internet access and computers), digital resources that offer meaningful value, and the skills to make use of them safely. Addressing digital divides could help ensure that AI development does not widen existing inequalities, and could also help respond to the long-term diversity crisis in the sector itself. Measures could include developing infrastructure in poorly-served areas and ensuring that existing infrastructure is used more effectively. Market measures could improve competitiveness and drive down prices, while regulatory measures could target minimum standards for ease of access and reliability. While

some households may never connect, free internet access in schools and public places such as town squares, libraries and other public buildings can fill the gap by providing access to computers and the internet. Several layers of digital skills and literacy initiatives could be introduced at all levels of the school system as well as adult education to promote everything from confident browsing to advanced programming.

- **Digital inclusion**. Technologies are increasingly customisable to serve individual needs and preferences, and people with disabilities are often enthusiastic early adopters of technology. While AI tools present opportunities to remove barriers to employment, education and social inclusion, there is also a risk they construct and enforce new criteria for 'normality' which are disabling for those that do not conform.[77] Just as replacing physical keypads with dynamic touchscreens can be disabling for the visually impaired, automatic driving applications that fail to recognise wheelchairs might create new mobility barriers for wheelchair users. Instead of requiring people to conform to their own definitions of normality, AI systems should respect the diversity of bodies, minds and personalities, and empower people to satisfy their own needs and preferences. To do so, AI developers could consult digital accessibility experts and apply universal design guidelines[78] that support the development of products for people with different abilities, that are flexible and accommodating to diverse preferences, simple and intuitive to use, present information appropriately, tolerate errors, require minimum effort and are physically accessible to all.
- **Beware 'reverse Turing' tests**. The Turing test examines whether a machine can be distinguished from humans as a means of defining whether or not it is intelligent. These tests have been reversed to examine whether a human can be distinguished from a machine. For example, CAPTCHAs[79] ask users to complete tasks that computers tend to fail in order to prove that they are human before accessing a service. As AI is increasingly deployed in roles that determine access to services – e.g. in moderating online content and filtering applications for jobs or loans – citizens are increasingly subjected to AI judgements of their trustworthiness, innocence or humanity. As human scrutiny of machines is mirrored by machine scrutiny of humans, it is crucial to ensure that decisions can be challenged in a timely manner, and that they are inclusive of all humans regardless of their level of ability.

4.3.5 Promote reflection and dialogue

In recent years, particularly since 2018, there has been an increasing level of debate about AI in policy, media, academic and cultural domains. This is beginning to shape the development and application of the technology. Several measures could promote deeper reflection and broader debate about how to manage AI in society in a range of settings.

Some AI-related challenges such as misinformation campaigns are particularly challenging for liberal Western democracies to counteract, as the most immediate solutions could interfere with European values such as freedom of expression and belief. However, these values and the actions they inspire

[77] M. Whittaker et al, Disability, bias, and AI, AI Now Institute, 2019.

[78] See European Disability Forum universal design principles.

[79] CAPTCHA stands for completely automated public Turing test to tell computers and humans apart. There are also increasingly common physical reverse Turing tests, which operate automatic lights by detecting movement in a room. Office workers regularly fail to meet the system's definition of humanity by not moving sufficiently, so the lights are switched off.

may form the most durable response to these challenges. Several of the measures set out in this study aim to double-down on liberal values through renewed focus on education, dialogue and social participation.

- **Harness the 'techlash'**. Concerns about democratic interference, fair taxation, privacy, 'fake news', employment and other issues have brought critical perspectives on digital technology to the mainstream. At the start of 2020, several observers heralded the year (or decade) of the 'techlash', that is, a backlash against the optimistic discourse and 'move fast and break things' approach to technology development. The significance of the techlash remains to be seen, but it could be harnessed into a constructive dialogue about what society expects and wants from digital technology, including AI, and what it is prepared to offer in return.
- **A digital social contract**. It may be possible to recast our relationship with digital tools in a digital 'social contract'[80] that is not formalised like a traditional contract, but rather sets out the roles, responsibilities and reasonable expectations of public authorities, private companies and individual users. Thinking about these trade-offs in the context of a digital social contract could support broad examination of the current state of affairs and reflection on how it should develop in the future. It could, for example, help approach the thorny issue of how media content – including news – is financed, evaluated and distributed, or help strike a balance between sovereignty and interdependence for citizens and states alike.
- **Understand and explain**. AI extends our ability to identify patterns and predict tendencies. While this can be useful, it does not always help us to understand underlying causal relationships and explain them in human terms, which are also important aspects of our knowledge about the world. As AI provides an abundance of information based on correlation, it is more important than ever to support other forms of knowledge production, including causation and contextualisation.
- **Reflection through culture and the arts**. Artistic output – most notably science fiction literature, cinema and television – has an important role in shaping public understandings of AI and its impacts.[81] However, the cultural sector can play a broader role in making sense of and re-imagining new technologies and our relationship with them. Culture and the arts could be promoted in this role by supporting work in certain areas as well as collaboration and skills development across the arts and computer sciences.[82]
- **Balanced media coverage**. Studies of media reporting on AI show that it is strongly influenced by industry hype regarding speculative opportunities. Academics account for very a small portion of news sources, and these are provided by a very small group of researchers that tend to have stronger industry links than citation records.[83] Encouraging wider consultation of

[80] Social contracts are broad tacit agreements about how individuals forego some freedoms and submit to certain authorities in exchange for the protection of rights and social order. They are constantly renegotiated and reinterpreted as the relationship between individuals and society develops, and can be subject to rapid change, for example as a result of revolutions and civil rights movements

[81] The word robot was coined by the playwright Capek. Other references to artistic output that regularly appear in AI debates include Asimov's laws of robotics, Huxley's *Brave New World*, Brooker's *Black Mirror* and, of course, Cameron's *Terminator*.

[82] See P. Boucher, Technology and the arts: Past, present and future synergies, 2019.

[83] See J. Brennen, A. Schulz, P. Howard and R. Nielsen, Industry, experts, or industry experts? Academic sourcing in news coverage of AI, Reuters Institute for the Study of Journalism, 2019.

scholars including those from a range of disciplines and ensuring that non-industry voices are represented could help support a productive and balanced debate on the issues.

- **Cross disciplinary education, training and research**. Cross-disciplinarity could be deployed in education, training and research to promote deeper reflection on AI in society. Technical elements can be embedded into social sciences and humanities curricula, as well as social and creative elements in computer science studies. Interdisciplinary training could be introduced in lifelong learning programmes for those that are in work, while funding schemes could be mobilised to encourage genuine cross-disciplinary collaboration in academic research.
- **Citizen participation**. A range of mechanisms facilitate deliberative and policy processes at EU level (including the European Citizens' Initiative and regular public consultations) and in Member States (such as the Citizens' Assembly in Ireland and the *'grand débat national'* in France). The forthcoming Conference on the Future of Europe is also anticipated to prioritise the inclusion of citizens' voices. These kinds of initiatives could be deployed to promote reflection and engagement on AI development.
- **Engage for public acceptability**. Public acceptance of AI has been cited as a key condition for the sector to flourish. Strategies that engage citizens early and produce technologies that are acceptable to them are more effective than those that produce technologies first and then find ways of encouraging people accept them. Indeed, beyond overcoming opposition, engagement can be used to generate active involvement and support to develop better technologies.[84]
- **Meaningful, informed dialogue**. Engagement processes are only meaningful when they substantially influence development paths. Consultations that do not have any effect risk damaging trust in both the technology and the actors that run the process. Also, if engagement is based upon imbalanced information, e.g. that is limited to applications for public services or overemphasising speculative risks, the results of the process will be of limited value. As such, dialogue should have a meaningful role in shaping development paths, and be informed by the full range of expected outcomes and uncertainties of development.

4.3.6 Refine the language

As discussed in Chapter 1, 'AI' is a problematic term because of both the ambiguous relativity of intelligence and the wide range of technologies, techniques, applications and contexts to which it refers. Using the same term to refer to autonomous weapons systems and medical diagnostic tools makes for needlessly confusing, divisive and unproductive debate. The quality of debate could be improved by refining the language that is used to talk about AI.

- **Not getting hung-up on intelligence**. Intelligence is a moving target. The Turing test – designed to set a threshold beyond which a machine would be considered intelligent – is regularly adjusted amid claims that today's AI may satisfy its conditions, but not its spirit. Passing such tests is a necessary but insufficient condition for intelligence. While it is normal for the goalposts to shift in response to developing capabilities and emerging possibilities, benchmarks such as error rates, impacts, and their distribution may be more relevant than intelligence.
- **Retire the term 'AI'**. Despite doubts about the appropriateness of 'intelligence' and the inclusivity of 'AI', AI remains the most common umbrella term for the full range of tools fitting its definition that are covered in this report, including expert systems, ML learning and several

[84] See P. Boucher, What if we could design better technologies through dialogue?, EPRS, European Parliament, 2019.

speculative future iterations. It is also likely to remain the most appropriate choice for public and non-specialist discussions. However, amongst specialists, the term 'AI' has become an obstacle to its own examination of the technology. It is so inclusive and ambiguous, and its use so opportunistic and unproductive, that many prefer to avoid it entirely in favour of more specific terms such as 'machine learning' which is, in most cases, what is under discussion.

- **Use specific terms where possible**. Arguments about specific impacts of AI rarely apply to all technologies and applications across the board. During debates, therefore, it is important to refer to them as specifically as possible, for example replacing 'AI' with 'machine learning tools for medical diagnostics', or choosing the most specific possible type of facial recognition (facial verification, identification or classification, see box in section 3.1.7). In some contexts, it can be important to highlight certain features, for example whether an application-sector combination leads to a definition of high-risk for liability purposes, or whether a service is powered in part by hidden human labour. It is also important to differentiate arguments about speculative future developments that may never occur from those about current AI that already affects society today, care could be taken to clearly qualify the former as 'speculative' or 'possible future' AI.
- **Develop a new language for AI**. The language of AI inspires parallels to human functions, most notably intelligence. In a way, this puts machines and humans in competition with each other. However, they have very different strengths and weaknesses. Perhaps subtle changes to the way we talk about AI could help us to imagine more complementary roles for humans and AI.
- **Choose metaphors carefully**. The AI domain is replete with metaphors such as 'intelligence', 'learning' and 'vision', which are similar but not quite the same as the human functions they are designed to simulate.[85] There are also several metaphors to describe AI development, so data becomes 'the new oil' and AI development is now a 'race'. The oil metaphor captures the value of data once it has been extracted and refined, but falls short on how data can be shared, reused and deleted. Describing AI development as a race provides an intuitive framing for global competition, but also implies that AI is a single technology, that there might be a single 'finish line' and a single winner. In reality, AI is a range of technologies and applications used by diverse actors operating in different contexts, with different goals and value systems. The race metaphor also fails to capture the role of cooperation, sharing and mutual benefits, which are key elements in the European approach to ensuring AI's quality and productivity. By framing discussions about global AI development as race, competition is prioritised over cooperation and participants feel compelled to follow the lead of those considered to be ahead. In this way, the race can quickly become a race to the bottom. As such, it is important to choose metaphors carefully, and reflect on what they imply before using them to frame discussions.

4.3.7 Choose appropriate applications and development paths

This final section presents measures that ground AI development paths in their features and capabilities.

- **Understand bias and subjectivity**. As explained in section 3.1.5, by training algorithms, we equip them with a worldview that they apply to individual cases. In this sense, they are discrimination machines. However, not all discrimination is alike. A good diagnostic support tool will discriminate consistently against images of cancerous cells. Bias that is based upon quality

[85] Sometimes these metaphors are reversed as brain functions are reduced to that of ANNs, repeating historic (mis)understandings of the brain as hydration systems, electronic fields and automatic rudders.

information about an individual, such as a credit score, differs from bias that is based upon statistical information about groups of people that are categorised as being similar. While embracing AI's discriminatory power, safeguards are needed to counteract the risk of reinforcing and exacerbating undesirable social bias and inequality. These could include a combination of technical, regulatory and social measures to better understand how algorithms make decisions, the impacts of these decisions including their distributional effects, and mechanisms for reporting problems and challenging decisions.

- **Avoid applications beyond AI's capabilities**. Some applications of AI are predicated upon concepts that we know to be false. For example, facial categorisation technologies that claim to be able to read emotions, identify sexuality, recognise mental health issues or predict performance. The problem is not only that AI cannot perform these tasks accurately, but that the suggested relationship between the input and output lacks scientific credibility and, like eugenics, could provide a baseless veneer of objectivity for biased decisions and structural inequalities. Similarly, ML algorithms are more suited to finding trends and correlations than causal relationships. This means they can be useful for making predictions where relationships are straightforward, but less so at predictions about individual social outcomes within complex systems. The application and level of autonomy granted to AI should be guided by a sound understanding of what is scientifically credible and within the capabilities of today's AI. This is particularly important in key decision-making domains such as employment, insurance and justice.
- **Avoid applications with undesirable impacts.** AI can be misused to predict performance from facial images or individual social outcomes from statistical data. The use of such tools can lead to unequal distribution of impacts and deviate from established principles such as the presumption of innocence.[86] Some AI applications that do perform well in their defined task can still be considered undesirable, such as personalised political advertisements in the context of election campaigns. The application and level of autonomy granted to AI should be guided by the understanding of factors beyond their direct aims, including their effectiveness and scientific basis, the wider impacts and their distribution, and their compatibility with social values. The development and use of algorithmic impact assessment could support this task.
- **Maintain human autonomy**. Despite some debate over the details, there is a broad consensus that humans should remain ultimately in control of AI. This may require some vigilance as the detection of automation bias have shown human propensities to accept the advice of automated machines over that of humans, occasionally with tragic consequences, such as aircraft crashes. As discussed in section 3.1.11, in contrast to the speculative challenge of resisting domination by intelligent machines, our greatest current challenge may be resisting over-reliance upon machines that are not as intelligent as we think. Various measures could be taken to counteract this vulnerability. In decision support systems, the human-machine interface could be set up to highlight or enforce the subordinate role and limitations of the AI. In consumer products – such as personal assistants and robot companions – measures could be taken to ensure the agent does not appear to be more intelligent or emotionally attached than they really are.
- **Look for solutions to problems, not problems for solutions**. While some AI developments are characterised by the development of solutions to identified problems, several more are

[86] Several examples are identified in C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown, 2016.

characterised by the search for problems to which existing methods could provide a profitable solution. While there is room for both models, the former might offer greater social value and more equitable distribution of benefits and risks. Mission-oriented research and public-private partnerships could be deployed to support the search for solutions to the problems we have (rather than problems for the solutions we have).

- **Consider what we really want from AI**. AI development and application is principally shaped by the availability of technology and data. Reorienting application decisions around desired change could help ensure that the impacts have a more profound social value that is more evenly distributed. This could be achieved by creating a stronger role for public dialogue and impact assessment in the articulation of current priorities as well as longer-term visions. For example, initiatives could be launched to assess and discuss how to finance the production and consumption of quality news in the digital age; how much public services should rely on private services such as maps; and how public debates should be moderated (e.g. via platform rules or public authorities).

5. Conclusions

This report presents how AI works and why it matters and sets out a range of options in response. Here, five recurring themes are brought together and presented as conclusions.

First, **language matters**. In many ways, the term 'AI' has become an obstacle to meaningful reflection and productive debate about the diverse range of technologies to which it refers. Meanwhile, debates about AI are replete with incomplete metaphors. Several of the options set out in the previous chapter target the way we talk about AI, including how we identify, understand and discuss specific technologies, as well as how we articulate visions of what we really want from AI.

Second, **algorithms are subjective**. AI learns from how society works, as represented in its data. From identifying pictures of cats to predicting music preferences, algorithms develop perspectives on the world which they use to make decisions. Since human societies have structural biases and inequalities, ML tools inevitably learn these too. While the only definitive solution to the problem is to remove bias and inequality from society, AI can only offer limited support for that mission. However, it is important to ensure that AI counteracts, rather than reinforces inequalities. There are some technical options to limit how algorithms can pick up on biases, although they might simply hide biases, making them more difficult to detect. Regulatory measures could limit how certain tools can be used in some domains, or provide recourse for when something goes wrong.

Third, **AI is not an end in itself**. The ultimate aim of supporting AI is not to maximise AI development per se, but to unlock some of the benefits that it promises to deliver. Thinking of AI more explicitly as a potential means of achieving other benefits could help maximise its social value. Instead of perfecting new technologies then searching for problems to which they could be a profitable solution, we could start by examining the problems we have and explore how AI could help us to find appropriate solutions. Similarly, public acceptability of AI could be sought by ensuring that the technology is acceptable to citizens, rather than encouraging them to accept it as it is. Greater trust in AI products and services could be fostered by designing more trustworthy systems, rather than encouraging citizens to have confidence in technologies that might let them down. In each case, meaningful dialogue with a range of stakeholders, including citizens, from the earliest stages of development could play a key role in defining what we aim to achieve, and how AI could help.

Fourth, **AI might fall short of its promises**. Many AI applications could offer profound social value. However, the best-known applications today offer relatively minor gains in efficiency or convenience, often with unevenly distributed costs and benefits. Employment impacts and privacy intrusions are increasingly tangible for citizens while the promised benefits to their health, wealth and environment remain intangible. On one hand, the promises could be moderated by making more modest claims about the capabilities of AI and the impacts of its application. On the other hand, more ambitious outcomes could be actively targeted by accelerating AI development that alleviates social biases and inequalities and provides direct and visible benefits for all.

Finally, **Europe needs to run its own AI race**. With major policy initiatives anticipated and preparations underway for the next generation of AI products and services based upon IoT and 5G, AI is at a pivotal moment for both regulation and technology development. The choices we make now could shape European life for decades to come. In running its own race, European AI can ensure a meaningful role for citizens to articulate what they expect from AI development and what they are ready to offer in return, to foster a competitive market that includes European SMEs, and to put adequate safeguards in place to align AI with European values and EU law.

Key references

European Parliament resolutions

Resolution of 1 June 2017 on digitising European industry, European Parliament.

Resolution of 16 February 2017 with recommendations to the Commission on civil law rules on robotics, European Parliament.

Resolution of 23 February 2018 on a European strategy on cooperative intelligent transport systems, European Parliament.

Resolution of 12 September 2018 on autonomous weapon systems, European Parliament.

Resolution of 15 January 2019 on autonomous driving in European transport, European Parliament.

Resolution of 12 February 2019 on a comprehensive European industrial policy on artificial intelligence and robotics, European Parliament.

Recommendation of 13 March 2019 to the Council and the Vice-President of the Commission / High Representative of the Union for Foreign Affairs and Security Policy concerning taking stock of the follow-up taken by the EEAS two years after the EP report on EU strategic communication to counteract propaganda against it by third parties, European Parliament.

Resolution of 12 February 2020 on automated decision-making processes: ensuring consumer protection and free movement of goods and services, European Parliament.

European Parliament publications

Kritikos, M., Artificial intelligence ante portas: Legal & ethical reflections, EPRS, European Parliament, 2019.

Madiega, T., EU guidelines on ethics in artificial intelligence: Context and implementation, EPRS, European Parliament, 2019.

Szczepański, M., Is data the new oil? Competition issues in the digital economy, EPRS, European Parliament, 2019.

Cesluk-Grajewski, M., Artificial intelligence: what think tanks are thinking, EPRS, European Parliament, 2020.

Should we fear artificial intelligence?, EPRS, European Parliament, 2018.

A governance framework for algorithmic accountability and transparency, EPRS, European Parliament, 2019.

Automated tackling of disinformation: Major challenges ahead, EPRS, European Parliament, 2019.

Cost of non-Europe in robotics and artificial intelligence, EPRS, European Parliament, 2019.

Understanding algorithmic decision-making: Opportunities and challenges, EPRS, European Parliament, 2019.

The ethics of artificial intelligence: Issues and initiatives, EPRS, European Parliament, 2020.

Ciucci, M., and Gouardères, F., The White Paper on Artificial Intelligence, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2020.

European Artificial Intelligence (AI) leadership, the path for an integrated vision, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2018.

Contribution to growth: The European Digital Single Market. Delivering economic benefits for citizens and businesses, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2019.

Education and employment of women in science, technology and the digital economy, including AI and its influence on gender equality, Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament, 2020.

Artificial intelligence (AI): new developments and innovations applied to e-commerce: Challenges to the functioning of the Internal Market, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2020.

New aspects and challenges in consumer protection: Digital services and artificial intelligence, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2020.

The e-commerce Directive as the cornerstone of the Internal Market, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2020

SME focus - Long term strategy for the European industrial future, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2020

European Commission communications

Communication on artificial intelligence for Europe, COM(2018) 237, European Commission, April 2018.

Communication on coordinated plan on artificial intelligence, COM(2018) 795 European Commission, December 2018.

Communication on building trust in human-centric artificial intelligence. COM(2019) 168, European Commission, April 2019.

Report on the safety and liability implications of artificial intelligence, the internet of things and robotics COM(2020) 64, European Commission, February 2020.

White paper on artificial intelligence – a European approach to excellence and trust, COM(2020) 65, European Commission, February 2020.

Communication on a European strategy for data. COM(2020) 66 European Commission, February 2020.

Other EU publications

Special Eurobarometer 460: Attitudes towards the impact of digitisation and automation on daily life, European Commission, 2017.

Shaping Europe's digital future, European Commission, 2020.

Statement on artificial intelligence, robotics and autonomous systems, European Group on Ethics in Science and New Technologies (EGE), 2018.

#BigData: Discrimination in data-supported decision making, European Union Agency for Fundamental Rights, 2018.

Expert Group on Liability and New Technologies – New Technologies Formation, Liability for artificial intelligence and other emerging digital technologies, European Commission, 2019.

High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, European Commission, 2019.

High-Level Expert Group on Artificial Intelligence, Trustworthy AI assessment list, European Commission, 2019.

High-Level Expert Group on Artificial Intelligence, Policy and investment recommendations for trustworthy AI European Commission, 2019.

Mazzucato, M. Mission-oriented research & innovation in the European Union: A problem-solving approach to fuel innovation-led growth., European Commission, 2018.

Samoili, S., M. Lopez Cobo, E. Gomez Gutierrez, G. De Prato, F. Martinez-Plumed, F. and B. Delipetrev Defining artificial intelligence, European Commission, 2020.

Sevoz, M. The future of work - work of the Future! On how artificial intelligence, robotics and automation are transforming jobs and the economy in Europe, European Commission 2019.

Books

Agrawal, A., Gans J., and Goldfarb, A., *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press, 2018.

Bostrom, N. *Superintelligence: Paths, dangers, strategies*, Oxford University Press, 2014.

Bridle, J., *New dark age: Technology and the end of the future*, Verso, 2018.

Collins, H., *Artifictional intelligence: Against humanity's surrender to computers*, Polity Press, 2018.

Fry, H. *Hello world: How to be human in the age of the machine*, Black Swan, 2019.

Harari, Y. N., *21 Lessons for the 21st Century*, Jonathan Cape, 2018.

Kurzweil, R. *How to create a mind: the secret of human thought revealed*, Duckworth, 2014.

Lee, K., *AI Superpowers: China, Silicon Valley, and the new world order*, Houghton Mifflin Harcourt, 2019.

O'Neil C., *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown, 2016.

Scharre, P., *Army of none: Autonomous weapons and the future of war*, W.W. Norton & Company, 2018.

Zuboff, S., *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books, 2019.

Academic journal articles

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. and Eckersley, P., Explainable machine learning in deployment, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*, Association for Computing Machinery, p648–657, 2020.

Flick, C., Informed consent and the Facebook emotional manipulation study, *Research Ethics*, 12(1), p14–28, 2016.

Kosinski, M., Stillwell D. and Graepel, T., Private traits and attributes are predictable from digital records of human behavior, *Proceedings of the National Academy of Sciences*, 10, p5802-5805, 2013.

Rosert, E. and Sauer, F., Prohibiting autonomous weapons: Put human dignity first, *Global Policy*, 10: p370-375, 2019.

Sparrow, R., The march of the robot dogs, *Ethics and Information Technology*, 4, p305–318, 2002.

Goddard, K., Roudsari A. and Wyatt, J., Automation bias: a systematic review of frequency, effect mediators, and mitigators, *Journal of the American Medical Informatics Association*, 19(1) p121–127, 2012.

Nemitz, P., Constitutional democracy and technology in the age of artificial intelligence, *Philosophical Transactions of the Royal Society*, 376 (2133), 2018.

Other reports

Brennen J., Schulz A., Howard P. and Nielsen R., Industry, experts, or industry experts? Academic sourcing in news coverage of AI, Reuters Institute for the Study of Journalism, 2019.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G.C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy R., and Amode D., The malicious use of artificial intelligence: forecasting, prevention, and mitigation, 2018.

European ethical charter on the use of artificial intelligence in judicial systems and their environment, European Commission for the Efficiency of Justice (CEPEJ), 2019.

Collins, A., Forged authenticity: Governing deepfake risks, EPFL International Risk Governance Center, 2019.

Pricing algorithms: Economic working paper on the use of algorithms to facilitate collusion and personalised pricing, Competition and Markets Authority, 2018.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, Nill Sánchez, A., Raji, D., Lisi Rankin, J., Richardson, R., Schultz, J., West, S.M., and Whittaker, M., AI Now 2019 Report, AI Now Institute, 2019.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, 2019.

Leslie, D., Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute, 2019.

Marzin, C., Plug and pray? A disability perspective on artificial intelligence, automated decision-making and emerging technologies, European Disability Forum, 2019.

West, S.M., Whittaker M., and Crawford, K., Discriminating systems: Gender, race and power in AI, AI Now Institute, 2019.

Whittaker M., Alper, M., Bennett, C., Hendren, S., Kaziunas, L., Mills, M., Ringel Morris M., Rankin, J., Rogers, E., Salas, M., and West, S.M., Disability, bias, and AI, AI Now Institute, 2019.

Yeung, K., A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework, Council of Europe. Committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT), 2019.

Artificial intelligence (AI) is probably the defining technology of the last decade, and perhaps also the next. The aim of this study is to support meaningful reflection and productive debate about AI by providing accessible information about the full range of current and speculative techniques and their associated impacts, and setting out a wide range of regulatory, technological and societal measures that could be mobilised in response.

OA-01-20-338-EN-N

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service

ISBN: 978-92-846-6770-3
doi: 10.2861/44572
QA-01-20-338-EN-N