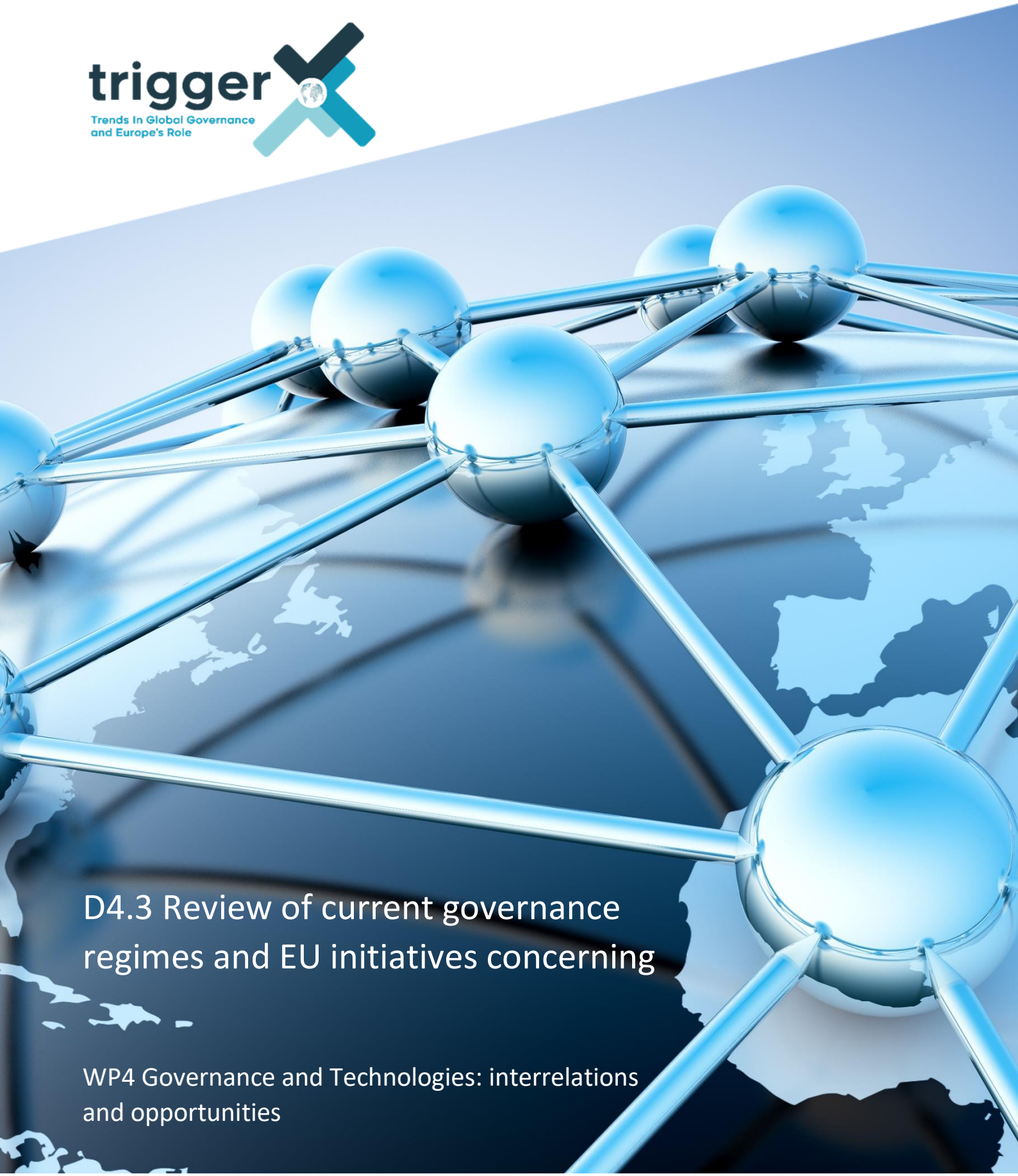
trigger
Trends In Global Governance and Europe's Role



D4.3 Review of current governance regimes and EU initiatives concerning

WP4 Governance and Technologies: interrelations and opportunities

Grant Agreement n° 822735, Research and Innovation Action

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement nº 822735. This document reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

TRIGGER

TRends in Global Governance and Europe's Role

Deliverable name:	D4.3 Review of current governance regimes and EU initiatives concerning AI (working paper)
WP / WP number:	WP4 Governance and Technologies: interrelations and opportunities
Delivery due date:	31.05.2020
Actual date of submission:	28.05.2020
Dissemination level:	Public
Lead beneficiary:	EPFL-IRGC
Contributors:	Aengus Collins, Marie-Valentine Florin, Anca Rusu, Anshuman Saxena, Gianluigi Viscusi
Reviewers:	Gianluca Misuraca, Andrea Renda

Changes with respect to the DoA

Not applicable

Dissemination and uptake

Public

Evidence of accomplishment

Report

Executive summary

This report is concerned with the governance implications of artificial intelligence and machine learning. Machine-learning models can be used to automate decisions in an ever-widening range of contexts and this automated decision-making can be deployed in ways that have a material impact on people's lives, from shaping the content of a social media newsfeed to determining whether or not someone should be incarcerated. As this phenomenon of automated decision-making increases in scope, so too does the need for processes to ensure the appropriate governance of these technologies.

In its simplest form, machine learning is the ability of a computer system to learn from input data and produce outcomes that can be used for diagnostics or predictions. There are two basic approaches: discriminative and generative. Broadly speaking, the discriminative approach learns how to distinguish between different classes that are present in a dataset, while the generative approach learns how to generate new samples that match the features present in a dataset. Applying these techniques to large datasets opens new possibilities in terms of analysing, predicting and even optimising human and institutional behaviour and decisions. This has led to machine learning being deployed across a rapidly widening range of sectors such as transportation, insurance, health and medical services, administration, media and advertising, and the military.

The use of machine learning offers numerous benefits, such as efficiency, scalability, analytical power and consistency, adaptability and convenience. However, it can also lead to serious consequences if things go wrong. With this in mind, it is possible to identify a number of overarching priorities that should frame the development of these technologies:

- Accuracy
- Bias
- Accountability and explainability
- Transparency
- Privacy
- Human oversight
- Broader ethical considerations

However, while these overarching priorities recur when machine learning is used, it is crucial to focus on how they interact in specific domains. It is only at the domain-specific level that the analysis of general patterns and principles can be sharpened into the identification of specific machine-learning challenges or problems that require concrete governance responses. In this report, we illustrate the importance of domain-specific factors by outlining how machine learning is used, and the different governance challenges is creates, in three sectors: autonomous vehicles, aspects of public administration, and healthcare.

As the sophistication and widespread use of machine learning has accelerated in recent years, the governance landscape has evolved in response. There has been a proliferation of initiatives and strategy documents aimed at establishing the kind of governance guardrails that will allow the potential benefits of machine learning to be exploited while mitigating potential risks. The governance landscape is dominated by the actions of three key players: the EU, the US and China. Accordingly, any recommendations as to the strategy that the EU should adopt in this area must take into account the actions of the other two.

Broadly speaking, a survey of the governance landscape identifies two main responses to machine learning: on the one hand, viewing it as a source of innovation, growth and competitiveness to be actively promoted, and on the other hand, viewing it as a source of potential societal disruption to be managed with caution. These are not mutually exclusive responses—on the contrary, the key factor is how different countries and regions seek to balance them. In general terms, the current situation is that the US and China are technologically and economically dominant in the machine-learning field, while the EU is more advanced in its governance approach, which is evolving towards a sophisticated regulatory framework designed to protect safety, consumer rights and fundamental rights.

This comparative strength of the EU in the area of governance—and in particular in the area of governance designed to safeguard fundamental values and rights—means that the EU should consider adopting an “ethics first” strategy of “niche leadership”, with a focus on those applications of machine learning where normative principles and ethical values play a particularly strong role. Of the three illustrative case domains that we cover in this report, this would mean prioritizing healthcare and public administration. The suggestion here is not the EU should depart the rest of the field, or stop pursuing innovation-led strategies. The argument is the more modest one that the EU is likely to find it more easy to exert influence in ethically charged domains, because it has already demonstrated credibility and effectiveness in projecting this kind of influence, notably in relation to data protection.

We suggest that there are four key recommendations that should shape the response of the EU to machine learning. First, it should develop its evolving governance framework (which already highlights the importance of focusing on high-risk sectors) by going beyond generic strategies for machine learning writ large, by focusing on concrete problems in specific domains. Second, it should build on its existing normative priorities and strengths, even if this means accepting the possibility that the innovation gap with the US and China will not be closed. Third, the EU should balance unity and diversity in its response. There is a lot to be said for the increased cohesion, authority and impact that comes from a coordinated EU approach to regulation, but in the context of a fast-evolving technology like machine learning, the potential benefits of regulatory experimentation across the member states should not be neglected. Fourth, and again with the pace of technological development in mind, the EU should seek to future-proof its governance

approach. This could be done using mechanisms such as “planned adaptive regulation”, which involves stipulating from the outset that regulations (and/or other governance mechanisms) will be revisited and amended over time on the basis of monitoring and feedback processes.

Contents

Index of Figures

Index of Tables

Index of Boxes

Introduction

The overall objective of the TRIGGER project is to develop knowledge and tools that the EU institutions will be able to use to enhance their role in global governance. Because digital technologies represent an increasingly important factor in the global governance landscape, they are an integral part of TRIGGER. The fourth work package examines the implications for the EU of a range of these technologies, which increasingly entail interconnected challenges related to governance of technology and governance by technology. This particular report is concerned with artificial intelligence. As per the EU High Level Group on AI (European Commisssion 2018a), the term "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals". In this report we focus more specifically on machine learning[1], where the challenges of "governance of and by technology" are particularly important. Machine-learning models can be used to automate decisions in an ever-widening range of contexts and this automated decision-making can be deployed in ways that have a material impact on people's lives, from shaping the content of a social media newsfeed to determining whether or not someone should be incarcerated. This ability of embedded machine-learning models to take decisions about humans is one way of defining "governance by" technology. As this phenomenon increases in scope, so too does the need for processes to ensure the appropriate "governance of" these technologies.

This has been recognised for some time. In 2016, for example, Doneda and Almeida (2016) noted that: "With algorithms' increased use to fulfil complex tasks comes the risk of algorithms' use for manipulation, biases, censorship, social discrimination, violations of privacy and property rights, and more. To address such risks, the process of algorithm governance should be considered." In the intervening years, machine learning has continued to progress rapidly, with a corresponding acceleration in the scope for embedded "governance by" technology. However, the "governance of" this technology has not recorded anything like the same rate of evolution. Machine learning is therefore a good example of the so-called technology "pacing problem" with which governments, regulators and other stakeholders are increasingly confronted. (Marchant, Allenby, Herkert, 2011) Not only is there a gap between technology and governance, but the gap is widening over time. This leads to growing risks of damaging disruption if things go wrong. The challenges are particularly acute in respect of machine learning, because of its huge versatility. It is a game-

[1] Increasingly, the phrase "artificial intelligence" is used so broadly and imprecisely as to be potentially misleading. As discussed in chapter 1, we focus in this report on the more clearly delineated set of technologies known as machine learning. However, because the use of "artificial intelligence" is so prevalent—including, for example, in most of the governance initiatives discussed in Chapter 4—it is a phrase that will recur throughout this report. We also make repeated use of the acronyms AI for artificial intelligence, ML for machine learning and AI/ML to refer to both.

changing or frame-setting technology, with the capacity to influence the trajectory of many other technology areas.

Because of this versatility of machine learning, one of the things we stress in this report is the importance of drilling down to consider the governance implications that arise when machine learning is used in specific domains. Many of the same issues will recur across sectors, but in different permutations and subject to different emphases and trade-offs. This is why we follow our discussion of overarching governance priorities in chapter 2 with an examination of the different ways machine learning is used and the different sets of particular governance priorities that arise in three illustrative domains.

As well as applying in individual domains where machine learning is used, governance trade-offs also exist at the macro level when broad priorities are being identified. In February 2020, the EU took a significant step towards the creation of a regulatory framework for artificial intelligence and machine learning, with the publication by the European Commission of the white paper *On Artificial Intelligence: A European Approach to Excellence and Trust*. (European Commission, 2020a) This builds on an earlier Commission Communication which listed three key objectives in this area: technological and industrial capacity, labour-market and educational inclusion, and creating a legal and ethical framework (European Commission, 2018d). However, it will not always be possible to advance these objectives simultaneously. There may be direct trade-offs between them. There is an analogy to be drawn here with the field of data protection where the GDPR, with its extraterritorial reach, has positioned the EU as a global "super-regulator" (Chander, Kaminski & McGeveran, 2020). However, for the most part this involves shaping the use of technologies that are developed and owned outside the EU.

Moreover, the equivalent governance trade-offs are more complicated for machine learning than for data protection. Data protection starts with a principle and approaches various technologies with that principle in mind. The governance of machine learning starts from the other end of the telescope—it begins with an exceptionally broad technology and tries to ascertain which principles, values or interests it affects. This is another reason for adopting a domain-specific approach as swiftly as possible—to shift from a broad question like "what should we do about machine learning?" to more tractable questions such as the following, all of which can be contextualised according to the various governance structures that already exist in the domains in question:

- "when assessing accident rates for autonomous vehicles, is it sufficient to match the rates for human-driven vehicles or is greater safety expected?"
- "in the public sector, which decisions can be automated without raising concerns about fairness, transparency and due-process?"
- "how do you regulate an algorithmic medical device that is going to evolve through numerous software versions over the course of a patient's life?"

This report proceeds as follows. Chapter 1 provides a concise overview of the “state of the art” in machine-learning technology. Chapter 2 sets out a number of overarching governance challenges, highlighting important normative considerations that need to be addressed in order to ensure the technology’s societal acceptability. In Chapter 3 we turn to consider how machine-learning is used in specific domains. Enumerating general principles and guidelines for machine learning is necessary but not sufficient. Governance of a technology as broad as machine learning requires taking account of the differing priorities, constraints and trade-offs that apply in various domains. We focus on three such domains: autonomous vehicles, public administration and healthcare. Chapter 4 surveys the current governance landscape for machine learning, with a particular focus on how the evolving approach of the EU compares with developments in the US and China. In addition to focusing on public-sector governance initiatives we also consider work being done on machine learning by universities, NGOs and scientific institutions.

In the remainder of the report, we build on the first four chapters to look more closely at the EU’s position in the evolving governance landscape. We highlight some of the opportunities, constraints and trade-offs that it faces, with a view to using this analysis to point the way to possible steps the EU might take to increase its influence on global governance in this area. In Chapter 5, we suggest three key themes that characterise the global governance landscape for machine-learning: (i) the imperative to capitalise on machine learning as a potential driver of economic growth; (ii) the need to address the kind of normative considerations discussed in Chapter 2; and (iii) the role of underlying societal values in balancing the first two. The chapter suggests a potential “niche leadership” position for the EU, which would see its governance influence concentrated in machine-learning domains where ethical constraints are particularly important. It looks at stakeholders' roles, opportunities and risks, and places the idea of EU “niche leadership” in the broader context of the influence over global governance that is currently enjoyed by a diverse range of actors both in Europe and globally. In Chapter 6 we consider the potential implications of a “niche leadership” role for the EU, suggesting ways in which the benefits of such a strategy could be maximized and the costs minimized. Chapter 7 concludes the report with a series of recommendations for the EU, and with a number of wider questions related to technology governance that inform subsequent tasks within the TRIGGER project.

1. Understanding machine learning

This chapter provides a brief overview of machine learning (ML) technologies, a sub-field of artificial intelligence (AI), with a view to facilitate an informed discussion on their governance or the risks that they generate. For this reason, the focus is both on introducing the fundamentals to support a basic understanding of the machine learning technology, and on highlighting practical issues related to the use of ML in production systems.

1.1. Overview

Artificial intelligence is a scientific discipline that resides at the intersection of philosophy, psychology and computer science, and aims to bring intelligent behaviour to machine-based problem-solving. The overarching motivation for AI has been to develop tools and apparatus capable of artificially synthesizing human-cognitive capabilities.

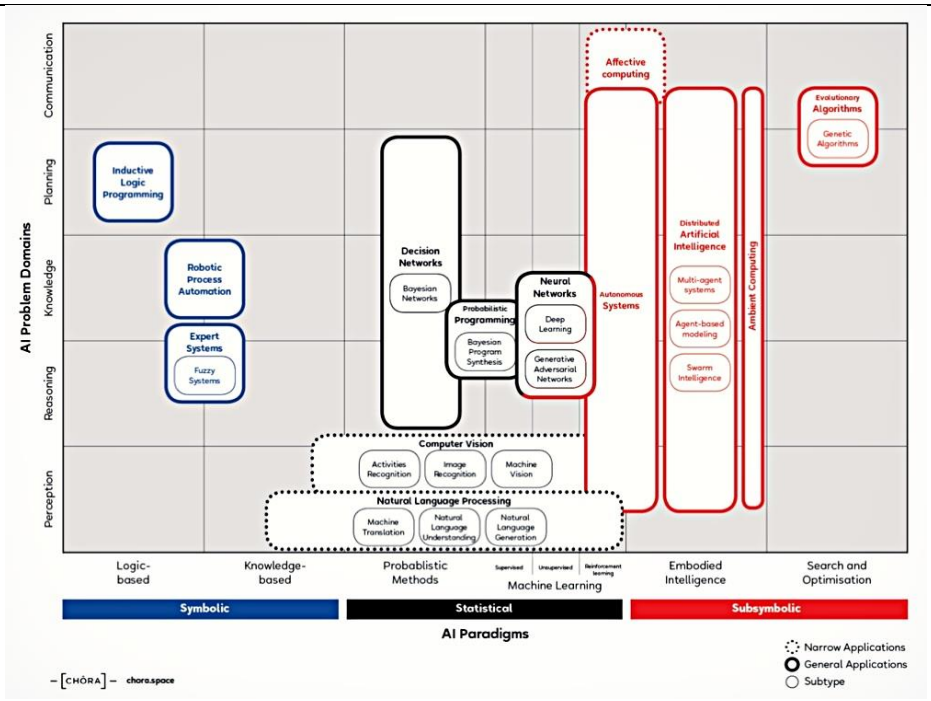


Illustration 1: The AI knowledge (Corea, 2018).

Machine Learning works at the intersection of the statistical (addressing big data challenges) and subsymbolic approach (addressing challenges to building learning systems) to AI

Over the years, the field of AI has expanded widely. The advancements in artificial intelligence are fuelled by several paradigms, each providing a distinct conceptual framework to realize machine-based problem-solving. The AI Knowledge Map in Illustration 1 presents the different AI technologies that engage these paradigms to realize basic cognitive capabilities such as perception, reasoning, knowledge, planning and communication. The term artificial intelligence is used increasingly broadly and imprecisely, in ways that can obscure the concrete governance

challenges that arise. In this report, we will focus more specifically on machine learning (ML), a subset of artificial intelligence that enables computer systems to learn from data and produce outcomes that can be used for diagnostics or predictions.

Machine-learning systems learn by extracting clues from raw-data using generic statistical constructs, such as smoothness, clustering, spatio-temporal coherence and sparsity, and by organizing these clues in neuro-biologically inspired computational architectures, such as the Neural Networks, to surface latent representations underlying the input data. In the following subsections, we discuss various aspects of ML technology with the view to facilitate the discussion on its governance.

1.2. Machine learning methods, types and applications

In its simplest form, machine learning is the ability of a computer system to learn latent representations from input data. There are two basic approaches to machine learning – the ***discriminative*** approach and the ***generative*** approach (Ng & Jordan, 2002).

The discriminative approach seeks to learn the difference between the classes present in the dataset. Learning is accomplished using labels that specify the class each data sample represents. The labels act as a supervisory signal that aid in learning the decision boundary that best separates these classes. Discriminative methods are particularly efficient at classification tasks as they learn the difference between the various classes present in the dataset and not what constitutes these classes. For example, while classifying animals, a discriminative method learns how one animal differs from the other and uses this information to classify previously unseen images of animals. On the other hand, the generative approach to ML seeks to learn the features that define the classes present in the dataset. As the name suggests, generative methods are most efficient at generating new instances (samples) of the classes identified during the learning process. When used for classification tasks, generative methods tend to be less efficient as the classification is based on establishing if a sample exhibits the features of a class not how different it is from all other classes.

The learning process can be facilitated by varying levels of human supervision, and accordingly, it may be categorised in one of the following four types - ***supervised***, ***self-supervised***, ***reinforced*** and ***unsupervised*** machine learning. Much depends on the task at hand and the input data available for training the ML model.

For classification tasks, which typically employs a discriminative approach, human-generated labels are used to inform the learning process about the class to which each training sample belongs. This type of learning is referred to as the ***supervised learning*** and is widely used in image classification and object detection. For example, AlexNet (Krizhevsky et al., 2012) is a popular image classifier trained on 1.2 million high-resolution images from the ImageNet dataset

(Deng et al., 2009) using purely supervised learning. The labels for these images were generated using the Human Intelligence Task (HIT) based service from Amazon Mechanical Turk.

For tasks such as autonomous driving, some aspects are learnt using inputs that are automatically collected. For example, (Bojarski, et al., 2016) demonstrate an end-to-end system that learns the control of steering-wheel using a dataset of road images – the training data, paired with the steering angles generated by a human driver. In this case, the steering angle is the supervisory input and the type of learning is ***self-supervised*** as the supervisory signal is not manually generated – it is automatically recorded without any manual assistance.

For tasks where systems actively interact with the surrounding environment present different opportunities for supervision. For example, while using ML for robotics, the learning model is supplied with outcome-related incentives to guide the learning process. The robot's space of interactions is profiled by assigning rewards for desired states and penalties for the undesired states. As the robot interacts with its surrounding, the incentives associated with the states reinforce the learning process and is thus referred to as ***reinforcement learning***. When compared to supervised learning, incentives are similar to labels - the difference lies in the timing of their use. In reinforcement learning, labels are associated with the outcome of an iteration, whereas in supervised learning labels are associated with each sample - many samples used in one iteration. In addition to robot design, reinforcement learning is also widely used for learning games. One example is AlphaGo, an ML model for the board game GO (Silver, 2017).

For other tasks such as art and other creative endeavours, the focus is on generating new instances of data as opposed to classifying existing data. These tasks employ Generative ML methods and focus on learning the representation underlying the training dataset. Learning the underlying representation with no supervisory input is referred to as ***unsupervised learning***. Unsupervised learning is often used as a pre-processing step to improve supervised learning results (Aisoma, 2018). Though much of its recent popularity is due to its use in conjunction with a classifier model that significantly improves the quality of newly generated samples. This combination of two learning models – the unsupervised model that generates new samples, and the supervised model that classifies newly generated samples as being new or not, is referred to as Generative Adversarial Network (GAN) (Goodfellow, et al., 2014). GANs have seen widespread adoption in image synthesis, where the quality of synthesized image has steadily improved during the last 5 years (see Illustration 2).

Illustration 2: Images of people who don't exist.

The quality of ML based image-synthesis has tremendously improved in the last few years,
Source: Twitter post by Ian Goodfellow, 14 Jan 2019

1.3. Model design

For most real-word AI problems—including speech and object recognition—the number of factors influencing the target phenomenon is very large. This quickly leads to huge levels of computational complexity, and one of the reasons for the explosion of interest in machine learning is its success/efficiency at dealing with such complexity.

For example, in the case of handwriting recognition, a 28x28 pixel grayscale image set is widely used for training ML models. Assuming variations are possible in any of the 784 pixels of the image-sample, a total of 2^{784} different image samples are possible. Mapping all the variations at the pixel-level directly to the set of visually discernible features at the image-level requires that each possible variation is present at least once in the training dataset. For the handwriting recognition example, this means a training dataset with 2^{784} image samples – an infeasibly large requirement. This is overcome by adopting a multi-level learning process where features learnt at the lower levels are used to compose more complex features at the higher level. This technique, called ***deep (machine) learning*****,** results in an accuracy of 99.65% for handwriting recognition tasks using merely 60,000 image samples training a 9 layered deep neural network (Ciresan et al., 2010). Illustration 3 provides a visual interpretation of how the features of an image are learnt using a deep machine learning model.

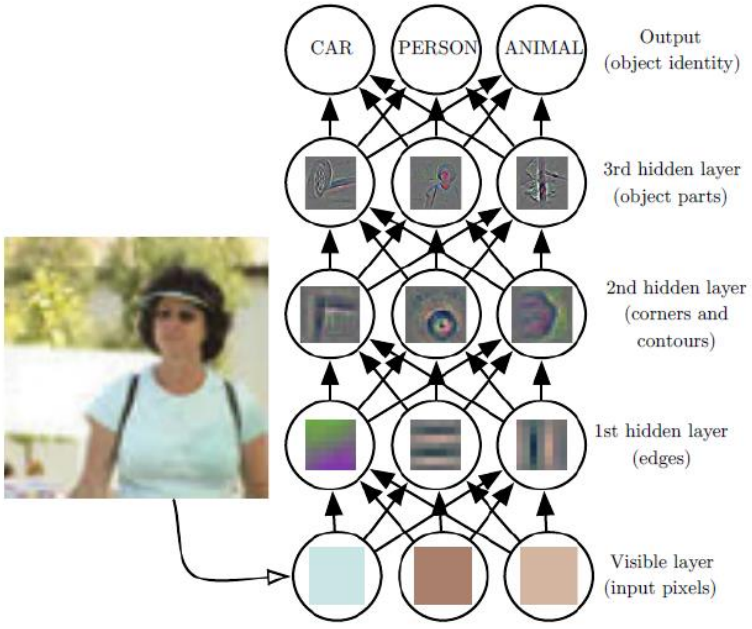


Illustration 3: Deep learning architectures

This visual depiction of how the feature set is built-up in a hierarchical manner in deep learning architectures. The learning starts by providing raw pixel values of the training image as input. At the first level the pixel values are used to learn simple shapes like edges, at the next level the edges are combined to learn more complex shapes such as corners and contours, at the third level the corners and contours are used to learning shapes of object parts and finally object parts are put together to recognize object identities.
Source: (Goodfellow et al., 2016; Zeiler & Fergus, 2014)

Deep Learning enhances the expressivity of machine learning models. ***Expressiveness*** is the ability of an ML model to learn complex representations from training data. The expressivity of neural networks increases exponentially with increase in depth (Zhang, Bengio, Hardt, Recht, & Vinyals, 2017). Nevertheless, adding more layers is beneficial only to a certain limit, beyond that, the model tends to overfit - the learnt representation fits the data too well capturing even the noisy and outlier elements that should otherwise be discarded. An ***overfit*** model tends to exhibit high variance as it mirrors the twist and turns of the training data and a low bias as it does not hold on to any initially specified shape. On the other hand, having less-than-adequate number of layers in the neural network leads to underfitting - the model is unable to learn the underlying relationships. An ***underfit*** model tends to exhibit high bias as it retains the initially specified shape, and low variance as it remains unaffected by the twists and turns of the training data.

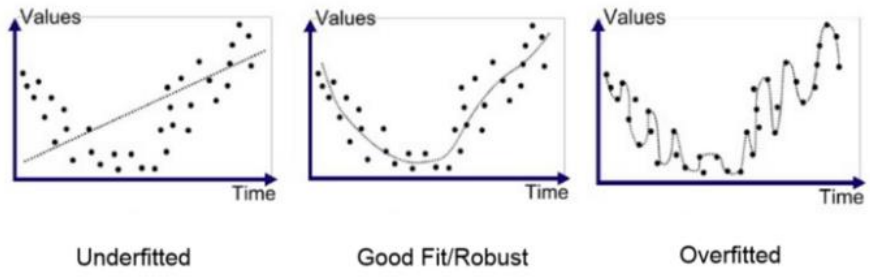


Illustration 4: A visual depiction of underfit and overfit as a curve-fitting problem.

Underfit (high bias, low variance) and overfit (low bias, high variance). The curve represents the approximation that the machine learning algorithm has learnt as a representation of the input data points.
Source: (Liew, 2019)

Both overfit and underfit models are unable to accurately predict or classify previously unseen samples. This is referred to as the ***generalization error*** and forms the primary basis for ascertaining the optimal depth of a deep learning model. Illustration 4 shows the underfit, overfit

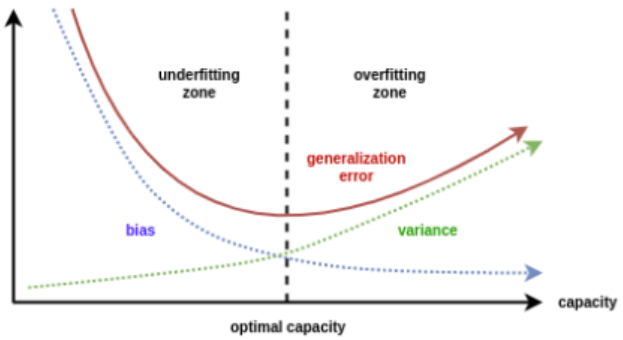


Illustration 5: Optimal model design as a generalization-error guided tradeoff.

As the depth of the neural network increases the model gains additional capacity which it uses to learn finer details from the data. Both, too-much and too-little capacity may lead to a model that simply mimics the data or completely discards the data. In either case, the learning if minimal and the use of such models on new datasets yield high error rates.
Source: (Saunders, 2017).

and good-fit situations as a curve-fitting problem, and Illustration 5 shows the model design trade-offs as a function of model capacity, i.e. the depth of the model.

Much of the effort in model design and training is empirically founded. Some best practices have emerged, though the field still lacks theoretical results that can provide strict performance guarantees and thus reduce the trial and error effort required to train an optimized model.

1.4. Privacy-preserving technologies

This report will emphasize the importance of privacy for the governance of AI/ML. For this purpose, this chapter concludes with a summary of the main techniques to ensure privacy when private data are stored on a server, processed and used.[2]

The field of **cryptography** provides several techniques to ensure privacy, integrity and authenticity of information exchange. Encryption, hash functions and digital signatures are the three basic tools available to system designers to build secure information systems. The field of encryption in particular has produced various methods, each of which aiming to distinct applications or objectives.

Digital privacy is a system property and cannot be achieved by simply subscribing to the use of cryptographic primitives. Much depends on the design of the security protocol – the set of rules that embed the cryptographic primitives in an application context, and the high-level system requirements that need to be serviced. At the system level, privacy shortcomings can be attributed to one or more of the following three factors: (a) Indifference: when a digital services provider purposefully engages in controversial practices to accumulate data until resistance is encountered; (b) Ignorance: when users willingly accept to share personal data in exchange for free services without evaluating the risks involved, and (c) Incompetence: when privacy-protection is inhibited due to lack of awareness about the risks or lack of access to tools, techniques and knowledge on responsible and informed use of digital services

More recently, the concept of **differential privacy** has been proposed as a guideline for designing privacy-protecting big data systems. Differential privacy is formal model of privacy that seeks to address the paradox of learning nothing about an individual while learning useful information about a population. Differential privacy algorithms provide privacy by introducing random noise in the process of information gathering thereby allowing plausible deniability of individual information. Nevertheless, differential privacy algorithms suffer from information leakage problem – each query on the dataset leaks some information and repeated use risks leaking user-data. Differential privacy systems thus require careful consideration in striking a fine balance between maximizing utility from data and preserving user-privacy.

[2] For a brief introduction to privacy-preserving technologies, see https://medium.com/@bertcmiller/emerging-privacy-preserving-technologies-are-game-changing-f32f06ac6aa and to privacy and data protection, see https://medium.com/better-programming/privacy-and-data-protection-c4f38678c639

Summary of technologies for privacy and security protection		
	Strengths	Weakenesses
Traditional encryption	Protects data at rest and in transit	Cannot protect computation
Homomorphic encryption	Protects computation in untrusted environment	Limited versatility vs efficiency
Secure multiparty computation	Protects computation in distributed environments	High communication overhead
Trusted execution environments	Protects computation with hardware trusted elements	Requires trust in the manufacturer, vulnerable to side-channels
Differential privacy	Protects released data from inferences	Degrades data utility (privacy-utility tradeoff)
Distributed ledger technologies (blockchains)	Strong accountability and traceability in distributed environments	Usually no data privacy

The EU GDPR refers to privacy by *design* and privacy by *default* to which one can also add privacy by *use*.

- Privacy by Design is an approach to protecting privacy by embedding it into the design specifications of information technologies, accountable business practices, and networked infrastructures, right from the outset. The following design considerations are particularly emphasised:
  a. Proactive not reactive - preventive not remedial
  b. Full functionality - positive-sum not zero-sum
  c. End-to-end security - full lifecycle protection
  d. Visibility and transparency - keep it open
  e. Respect for user privacy - keep it user-centric
- Privacy by Default is aimed at promoting the practice of provisioning digital systems with a default setting of highest privacy supported. This is to safeguard an ignorant user who is not well-informed about the various privacy settings when he first starts using a digital product or service. He might continue the default settings for a while before taking full command of the system configurations.
- Privacy by Use is a reminder that privacy ultimately depends on how much the user values privacy and to what extent he/she is diligent in his/her digital behavior. And for this reason, digital literacy and privacy education remain very relevant.

One method of privacy preservation with particular relevance for machine learning is **federated (or distributed) learning.** This enables several actors to build a common model without pooling data. Instead, the ML algorithms are trained on multiple local datasets. In principle, this addresses critical issues such as data privacy, data security, data access rights and access to heterogeneous data.

2. Benefits, risks and governance

The machine-learning technologies described in the preceding chapter are becoming increasingly deeply enmeshed in the functioning of societies and economies. The promise of leveraging big data to analyse, predict and ultimately optimise human and institutional decisions has seen the deployment of machine learning widen rapidly across sectors such as transportation, insurance, health and medical services, administration, media and advertising, and the military.

The rapid development and deployment of machine learning presents huge challenges, particularly when decision-making algorithms are used in contexts that require interpretation, judgment and ethical deliberation, and especially when humans are no longer in control. In this chapter, we draw on the insights of interdisciplinary experts to highlight a number of dimensions of machine learning that require particular attention from a governance perspective.[3]

2.1. Benefits and risks: an overview

The use of machine learning—particularly if decision-making is delegated to algorithms—can lead to potentially serious consequences if things go wrong, such as regulatory infringement, for example, or a contractual breach. Prior to delegating a decision to a machine, therefore, it is crucial to understand the potential benefits and risks, as well as the various trade-offs that will inevitably exist between them.

Potential Benefits

- **Efficiency:** generating outcomes more rapidly and at a lower cost
- **Analytical power:** processing huge and diverse datasets, finding patterns and linkages in ways not possible for humans
- **Scalability:** operating across vast domains—geographic, subject-matter, or otherwise
- **Consistency:** processing information reliably and systematically regardless of scale
- **Adaptability:** rapid (even real-time) algorithmic learning in response to changing data
- **Convenience:** performing tedious or time-consuming tasks, freeing up human time for more meaningful or higher-value pursuits

Potential Risks

- **Errors:** difficult to identify or correct due to the opacity of many algorithms (the “black box” problem). However, software correctness has always been a problem, and perfect correctness does not exist except in some domains such as aviation, where software are not built to evolve or update.

[3] This chapter draws on the proceedings of an EPFL International Risk Governance Center expert workshop entitled “The Governance of Decision Making Algorithms”. See EPFL IRGC (2018). The Governance of Decision-Making Algorithms. Lausanne: EPFL International Risk Governance Center.

- **Software bugs:** increasing complexity can make is impossible to test the "correctness" of software comprehensively
- **Data breaches:** tension between privacy rights machine-learning's need for large datasets
- **Bias:** the algorithmic learning process can entrench discrimination related to race, gender, ethnicity, age, income, etc
- **Loss of human oversight:** algorithms are now capable of making decisions in domains where judgement and wisdom have always been thought essential
- **Unaccountability:** difficult to assess responsibilities or liabilities relating to black-box decisions
- **Social control:** using machine learning to surveil citizens, influence behaviour or interfere in elections
- **Criminality:** leveraging machine learning to make rule-breaking more efficient, scalable, powerful, etc

Note that the risks described above overlap. Some are at the technical level, others at the societal level. Views on risk depend on perspectives and who is concerned. There is a hierarchy of risks that manifest at various levels (scales) with, for example, data breaches possibly leading to criminality (and vice-versa), and loss of human oversight connected to, although different from social control.

It is important to note that the risks and benefits of algorithmic decision-making need to be considered and weighed not in the abstract, but in concrete domain-specific applications, where important distinctions may emerge. The table below illustrates a number of such domain-specific examples. In Chapter 3 we expand on this by providing a more detailed discussion of the way machine learning is used, and the governance challenges that arise, in three domains: autonomous vehicles, public administration and healthcare.

Table 1: Examples of domain-specific risks and benefits

Domain	Potential risks	Expected benefits
Automated driving	Wrong assessment of a car environment (car-to-car and car-to- infrastructure) leading to an accident	Benefits of autonomous (connected) guiding of vehicles, such as increased traffic efficiency and fewer accidents Comfort and convenience
Public administration • **Criminal justice**		Ability to enforce rules a priori by embedding them into code

• **Public services / social benefits** • **Facial recognition**	Incorrect prediction of recidivism, potential unfair discrimination Incorrect, potentially unfair discriminative distribution of social benefits Undue or illegal citizen surveillance	Embedding into code rules for a loan or social benefit attribution Reducing eyewitness misidentification (a lead cause in wrongful convictions)
Medical diagnostics and prognostics	Wrong medical diagnosis, prognostic or treatment decision	Improving the capacity to diagnose, prevent or treat life-threatening diseases
Insurance contracts	Incorrect actuarial analysis misprices risk or introduces unfair discrimination in prices	More efficient allocation of risk, e.g. through better actuarial analysis and fraud detection

2.2. Overarching governance priorities

As the computational sophistication of machine-learning algorithms has advanced and their deployment has widened, the machine learning governance landscape has expanded and become increasingly crowded. (In chapter 4, we survey a selection of the numerous governance initiatives that have proliferated in recent years.) However, despite this growing complexity, it is possible to identify a number of overarching priorities that recur as being central to the development of these technologies. In this section, with a view to framing the discussion that follows in the rest of this report, we outline six of these priorities.

Accuracy

In order to tackle inaccuracies that may yield errors, bad decisions or misrepresentations, it is crucial to improve the quality and representativeness of data.[4] The success of a machine-learning system relies on access to large volumes of good quality data. Some of the aspects that are helpful in ascertaining the quality of a dataset include labelling errors, noisy features, duplication of examples and omitted values (Google, 2019). A low-quality dataset requires significant pre-processing before it can be used for training purposes. By some estimates, data scientists spend around 80% of their time on preparing and managing data for analysis (Press, 2016). As a result, access to trusted and well-curated datasets remains a top priority for the ML research community. For specialized tasks, such as imaging for medical diagnostics, there may be additional concerns of privacy and ownership that may require a concerted effort by both public and private players.

[4] The lack of adequate training data can undermine the potential of machine-learning systems, as this assessment from the healthcare sector attests: "If Watson has not, as of yet, accomplished a great deal along those lines, one big reason is that it needs certain types of data to be 'trained'. And in many cases such data is in very short supply or difficult to access. That's not a problem unique to Watson. It's a catch-22 facing the entire field of machine learning for health care." (Freedman, 2017)

As an illustration, automated driving will not improve on safety and reliability if it lacks sufficient training data, if it does not pick up on common human mistakes (such as not respecting stop signs), or if the training is concentrated in one type of setting or with an insufficiently varied population of drivers. When machine learning is used in the criminal justice system, for example to predict recidivism, similar risks of inaccurate results are created if algorithms are not exposed to a sufficiently broad dataset[5].

However, there is an important qualification to introduce here. While data limitations can be a source of inaccuracies and consequent problems, this does not mean that the answer is to rely on ever-larger datasets. For one thing, as noted before, data quality matters as well as quantity. More fundamentally, however, relying on increased levels of datafication can obscure the range and nature of the problems that need to be addressed. The trend to quantification may involve a loss in meaning, interpretation, nuance and a rich understanding of social reality, where human judgement matters.

Moreover, there are important societal reasons for limiting the availability of certain data. For example, removing or anonymising certain data may potentially lead to a less "accurate" dataset for algorithmic learning to work with, but may be crucial to to upholding users' fundamental rights, notably in relation to privacy, consent and similar principles. The existence of a potential trade-off between accuracy and data protection highlights the importance of the privacy-preserving principles and technologies discussed in Chapter 1 (section 1.4).

Another potential drag on the accuracy of machine learning systems relates to the diversity of the data used. In the process of learning and optimizing, algorithms may exclude information deemed irrelevant, but this process could lead to a reduction in the diversity of information that users and operators actually encounter, which in turn can lead to suboptimal collective decisions or behaviour[6]. Steps can be taken to mitigate this risk in some areas, but one fixed constraint of machine learning is its reliance on data relating to the past. This has potentially serious societal consequences: a model may be highly accurate when judged against the past, but may transpire to be much less accurate if deployed in changed circumstances. The digital transformation of society requires not only accurate technological solutions but, as Floridi among others has argued, a wider rethinking of important aspects of how society functions. (Floridi, 2015)

Accuracy also requires some interpretive nuance. What is "accurate" for one group of the population might not be for another, but these distinctions may be lost in the aggregate picture that the algorithm develops. This has been noted in relation to facial recognition technologies (Reilly, 2018), but it also matters in other contexts such as medical diagnosis and treatment,

[5] On the methodological intricacies of comparing bail decisions of judges to machine-learned algorithms of what a defendant would do if released, see Kleinberg et al. (2017).

[6] This is familiar from analysis of "herd" dynamics in financial markets. In a machine-learning context, if all oil-trading companies trained their trading algorithms on the same datasets, it would create the risk of all market players selling or buying the same assets at the same time, potentially leading to dangerous market disruptions.

attribution of criminality, and automated driving. Accuracy also needs to be considered in disaggregated terms. For example, an algorithm using past healthcare claims to predict users' number of hospital visits may have high accuracy overall, but without proper contextualization in interpreting such results, healthcare providers or insurers may unduly target the elderly or other sub-populations with certain pre-existing conditions (Tramèr et al., 2017).

Linked to a broader definition of accuracy, O'Neil has shown how using complex algorithmic computations to attribute students' educational results to their teachers can yield suboptimal, even unfair outcomes such as dismissing capable teachers and/or incentivizing teachers to game the system rather than prioritize students' needs (O'Neil, 2016). In this example, the problem is broader than simply accuracy: it is a problem of social adaption to the introduction of measurement instruments. Once ML algorithms get used to govern, the rules of the game change and actors will change their behaviour accordingly (including ways to "game the system"). This illustrates that the introduction of ML in any kind of environment (especially if it is done for management purposes) will trigger new responses.

An overarching concern is how to evaluate machine learning's accuracy against the human baseline. Algorithms may have the potential to reduce the quantity of errors, but those errors may be qualitatively worse when judged against the results of human decision-making. Another challenge is that we do not always know how to test machine learning algorithms. It thus becomes necessary to decide, perhaps even regulate, how we determine accuracy for algorithms.

Ultimately, one must recognise human fallibility on the one hand and the impossibility of error-free algorithmic decision-making on the other. Algorithms learn and process information in ways that differ from humans. They err differently too, and this risks leading to the rejection of potentially promising techniques and advances at the first failure.

While neither type of decision-making—human or algorithmic—is error-free, it may be possible to engineer safeguards in a decision-making system. Designing for potential failure—engineering for "better worst-case" algorithmic behaviours—might be as important as improving overall behaviour. Setting certain "guard-rails" or boundfaries which algorithms cannot cross—and demanding the ability to prove those boundaries as a matter of technical design—can help render algorithms more reliable over time.

Bias

Critical challenges related to data quality and algorithmic decision-making revolve around the problem of algorithmic bias. Biases can manifest themselves at different points or aspects, most notably in the data itself (directly or indirectly by way of proxies[7]), the manner in which they are

[7] Generally speaking, a proxy is someone or something representing another. In the case of algorithmic decision-making, it has been defined as a "feature correlated with a protected class whose use in a decision procedure can result in indirect discrimination" (Datta, Fredrikson, Ko, Mardiziel, & Sen, 2017).

processed/optimised, and also at the level of interpretation. Biases in decision-making are not new. Abundant evidence suggests that humans are often biased, which affects their decisions. One of the many questions around biases is whether humans are still better to interpret and correct their private and institutional biases, or whether computational solutions could be developed that proactively detect and correct problematic biases over time.

While not all bias is 'bad' per se, biased representations in data, learning context/method and outputs can lead to social discrimination. For example, the US Department of Housing and Development has filed a complaint that advertising algorithms on Facebook violate the Fair Housing Act as they discriminate against protected groups on the basis of disability, gender, race, ethnicity, age, etc. (Booker, 2018). What is further troubling about algorithmic bias is that when it facilitates undue discrimination, it may be systematic but not easily detected. For example, if biased machine-learning algorithms process the bulk of resumes or decide which jobs to advertise to whom, they may unintentionally discriminate against applicants but the users themselves may not be aware that they are being discriminated against (Carpenter, 2015).

De-biasing techniques exist, but there is an almost paradoxical situation at play: in order to evaluate whether these proxy measures creep into one's decision-making system, one would need more information and may have to include the very sensitive categories (such as protected categories of personal data) to know if one has sufficiently minimised the bias. Is this socially acceptable and does it require a legal or doctrinal shift?

There is an onus on machine-learning developers to use programming best practices for developing bias-free algorithms. A telling example is the case of a recruitment tool developed by the online retailer Amazon (Brookings, 2019). Programmers at Amazon used data derived from the resumes submitted over a 10 year period, which were predominantly from white males. The algorithm was taught to recognize word patterns in the resumes and the data was benchmarked against the predominantly male engineering team. As a result, any new submission from a candidate mentioning the word "women", such as having studied at a women's college, was automatically downgraded. Programmers, particularly those working on ML projects, can play an important role in mitigating the adverse effect of historical biases that will inadvertently be present in most datasets. As in the example above, the use of gender-free references or the use of skill-sets as opposed to human-identities could be one approach to avoid gender bias.

Accountability and explainability

Accountability is a broad concept that can be linked to other notions such as transparency, due process, fairness, or legal responsibility. It is also particularly interlinked with explainability – the degree to which it is possible to explain how decisions are made. This matters for understanding the provenance of important decisions and for enabling redress in case of erroneous decisions. As we will see, explaining algorithmic decisions can be challenging owing to algorithms' black-

box properties. However, the comparison with human decision-making is potentially instructive—it is not always the case that humans can explain their decisions in a regular, complete, consistent or accurate way.

An important part of the accountability discussion is the question: "accountable to whom?" The paragraphs that follow highlight some of the technological considerations around accountability, explainabilty and transparency. But the intention here is not to reduce the question of accountability to technical fixes. All of these issues must be considered in the context of the wider governance structures that are needed to ensure that accountability is robust and consistent: laws, norms, incentives, institutions, etc. This touches on one of the guiding questions of this report: to what extent can (and should) the European Union play a greater role in shaping these governance structures, thereby becoming an increasingly important guarantor of accountability. As we discuss in Chapter 4, the General Data Protection Regulation (GDPR) has begun the process of codifying EU-wide standards on the accountability and transparency of ML algorithms. (Officialblogunio, 2019) Work is also under way to review the EU's legal framework for product liability to ensure its applicability to new technologies. However, it is also worth noting that the enforcement of accountability does not always rely on regulatory intervention. Accountabiliy is often treated as a matter of reputation, with institutions accountable to the individuals affected by their ML systems: if the institution misbehaves, that will affect its reputation, possibly causing financial losses or reduced confidence and trust.

The opaque nature of machine learning such as neural network can limit audit and compliance capabilities. Given the strong data-dependent nature of machine learning, even small changes in model configurations can have unexpected consequences in the outcomes. Once an algorithm is deployed, additional challenges are introduced due to complex topologies with provision for retraining, inference-making, model approvals, model rollbacks all occurring simultaneously in real-time across a heterogeneous set of distributed environments (Sridhar et al., 2018).

Explainability can be particularly helpful in novel contexts: when a process is new or in flux, having some form of explainability can help to establish a baseline or confidence and understanding about how important decisions are made and about the reduction of wrongful or erroneous decisions.

Trade-offs can arise between pursuing explainability—for example, with reporting mechanisms or by restricting the use of machine learning so that only easily interpretable models are permitted—and pursuing computational prowess through more fluid or fast-evolving decision-making systems[8]. When explainability and performance have an inverse relationship, then assessing the impact on end-users and distinguishing high-stakes decisions from low-stakes ones may help

[8] In general, 'AI systems do not automatically store information about their decisions. Often, this feature is considered an advantage: unlike human decision-makers, AI system scan delete information to optimize their data storage and protect privacy' (Doshi-Velez et al., 2017)

define the degree and mode of explainability that is needed. For example, if algorithmic decision-making leads to a sustained positive impact on the life or livelihood of people, then with good performance, explainability becomes less of a requirement over time. For negative but significant outcomes, however, sacrificing performance for higher explainability is more salient. The underlying norm is that people should be able to know where a decision affecting them is coming from. The explanation may at times be more 'local' (e.g. how a specific decision was arrived at) or 'global' (e.g. how various factors are weighed and interlinked to produce a type/class of decisions). It may also be counterfactual, for example attempting to isolate the factor that made a difference between obtaining a loan or not[9]. Crucially, the explanation must be intelligible from an end users' perspective (i.e., it does not suffice to simply put some code in the open).

The following factors should be addressed when a machine learning system is being used:

- provenance: the ability to identify the exact sequence of events (datasets, trainings, code, pipelines, human approvals) that led to an outcome or event
- reproducibility: the ability to replay the above sequence and replicate the algorithmic prediction, thereby setting the context for investigating alternatives
- interpretability: the ability to explain what a model has learnt in human-understandable terms (Doshi-Velez & Kim, 2017)

Other concerns related to explainability include the possibility of simply mining a "conveniently plausible" explanation (e.g. 'adversarial explainability', p-hacking), as well as the challenge that feedback often benefits an adversary (i.e. any feedback from an algorithm might help those who are attacking it).

Transparency

Transparency is closely related to explainability. In this context, transparency does not mean transparence of the ML process, but applies primarily to organisations' choice of underlying algorithmic methodology as well as their willingness to respond to queries about outcomes in an open and accessible way (such as by making their decisions verifiable or accessible for independent assessments). Insofar as the digital space is perceived as dominated by a handful of players (notably technology giants and certain governments) with privileged access to data and computational know-how, a lack of transparency as to how such organisations curate the data they use for specific goals can also render any explanations of algorithmic outcomes suspect**.**

Along with some explainability of outcomes, additional layers of transparency (about the data input, learning process but also when decision-making algorithms are legally deployed) remain important for rendering algorithmic decision-making more accountable. While it resonates at a general level, the quest for transparency does not seem to be that straightforward as it plays out

[9] Or 'local counterfactual faithfulness' (Doshi-Velez et al., 2017)

in many dimensions. There may exist socially accepted reasons and legal grounds for why users or organisations care to protect private data, proprietary code, etc. Disclosures about access to private data and introducing more independent monitoring platforms to verify or certify the quality of the underlying data (e.g. assess legality, accuracy, etc.) can help maintain a more trusted relationship with the data subjects. At the same time, further clarity on when there is a 'right to explanation' and a right not to be subject to automated decision-making (see the discussion of the General Data Protection Regulation (GDPR) in Chapter 4) can also partially address the public's need to better understand the uses and limitations of machine learning.

There is on-going research about developing and enabling explainability and redress in case of failure, designing guardrails that algorithms cannot theoretically trespass as well as enabling human oversight in critical contexts. This is necessary to make them more socially acceptable.

Privacy

Privacy is one of the most prominent governance issues across most digital technologies, including machine learning. As we will see in chapter 4, it is of particular importance in Europe, where the EU's GDPR has established a sophisticated framework of protections for individuals' personal data. Such regulation is often 'principles-based', backed by more prescriptive rule-sets to embed those privacy-protection principles into law. Privacy is now one of the first criteria that needs to be considered when new technologies and developed and deployed. For example, at the time of writing (May 2020) debate about rapidly developed COVID-19 contact tracing apps has focused to a large degree on concerns that the data collected could be used against the interest of the individual.[10] Even though principles-based regulation is often more adaptable to new technology, the particular case of algorithmic decision-making is challenging.

The GDPR contains principles to ensure that the processing of personal data adequately matches its stated purposes ('purpose limitation') as disclosed to the data subject ('transparency'). These purposes must typically be recognised and stated to data subjects upfront. However, especially where decision-making algorithms are used, the purposes for which the decisions are made may change as a function of the learning of the algorithm, for example where decision-making algorithms are able to highlight hitherto opaque patterns. This may be by design of the technology, or more likely by an act of the operator, who may extend the use of the decision-making algorithms beyond their originally stated purposes. Without clear governance and controls, there is a risk that unchecked extension of the purposes of algorithmic decision-making—where personal data are being processed—could cause a breach of GDPR.

[10] The contact-tracing process itself does not rely on machine learning, but depending on the design of the system, a tracing app could generate huge datasets that could be used as inputs for ML systems. One possible use of this data would be for epidemiological purposes, to recognize patterns and make predictions about the spread of COVID-19. However, privacy advocates have stressed the need to collect only anonymous and aggregated data for epidemiological use, to prevent against new datasets being misused. In this debate it is difficult to disentangle the various aspects related to technology, to policy and regulation, and to public opinion.

Data minimisation is an important consideration in relation to privacy. This principle requires that the use of personal data be adequate, relevant and limited to what is necessary for the relevant purposes of processing. However, machine learning generally performs better with larger datasets from which algorithmic insights can be inferred. Does that mean that all such data are relevant and necessary? Or is there a threshold of relevance and necessity, below which data are not retained? If the latter, how does one apply such a threshold given that the relevance and necessity of particular data may become apparent only in hindsight—after the algorithm has observed enough data to be able to learn. This may present a risk both to privacy and to the adoption of algorithmic decision-making as organisations grapple with how to comply with evolving regulatory obligations. The use of appropriate cryptography techniques (or other privacy-preserving techniques) is another important consideration for protecting data and privacy. For example, techniques that enable to work on encrypted data on in *trusted hardware environments* are particularly promising.

Human oversight

In general, the more “independent” of human oversight and involvement a machine-learning algorithm is, the greater is the need for scrutiny. In addition to considering the sociological and economic aspects of this question (such as concerns as to whether algorithms will replace humans in societally destabilising ways), it is also important to clarify our expectations for how human and algorithmic decisions will relate or interact. When do we want or expect algorithmic decision making to mimic human judgment? When to aid it, correct it, or perhaps even override it? And when do we expect instead that humans will to be able to override algorithms?

One way to frame the question of human control vis-a-vis algorithmic decision-making is by clarifying in each instance whether humans**:**

- Are fully in control ('in the loop'), in the sense that, at some stage in the decision-making process, the algorithm stops, hands the decision over to a human, who then instructs the algorithm to proceed in a certain way;
- Can take control if needed ('on the loop'), in the sense that the algorithm informs a human supervisor who can intervene to modify the decision or to take control; and
- Cannot take control ('out of the loop'), either because there is no supervision by design (in hypothetical fully automated self-driving cars), or because decisions are irreversible (launching nuclear missiles).

This is not a precise taxonomy ; it is rather a heuristic to help evaluate when, how and with what latitude we can reasonably allow algorithms to take on decision-making attributes, and who assumes responsibility (especially legal, but moral too) in case of decisions that lead to harm.

In practice, many systems where humans should be able to take control actually put them out of control because of the sheer volume of data and decisions being produced by the algorithm,

making human supervision all but impossible. So while “can take control” may seem like the most balanced ‘default’ option, it comes with some risk that humans may struggle if control is abruptly handed back to them, because they lack the relevant context, practice, attention or time for making critical decisions under pressure. The evolution of autonomous vehicles provides numerous examples of accidents occurring because the driver cannot take control back. Similarly, in modern aircrafts it can be very difficult for pilots to take control back from autopilot mode.

Ethics

While each of the six priorities discussed above have an rethical or normative component, it is important to note that one of the growing governance challenges for machine learning is the fact that it can increasingly easily be deployed in what we might term ethically-sensitive or values-laden domains. This raises particular risks and therefore calls for increased vigilance. While algorithms may be able to learn continuously and perform certain tasks more quickly, consistently and accurately than humans, they have important limitations. They may make decisions that strike us as lacking common sense. They may miss important contextual changes. It is difficult for them to recognize their own errors, or to weigh the consequences of their decisions, or to be reflexive about societal preferences.

Ethical considerations have shaped the debate about many technologies in the past. However, while questions about the ethical implications *of technology* may apply across numerous technologies, arguably algorithmic decision-making sharpens dilemmas about the embedding of ethical values *in technology*. An algorithm’s lack of interpretive skills, common sense, self-awareness, empathy or social intelligence raises questions as to the type of decisions that can or should be delegated to it, and subject to what human supervision. Among other things there are important technical considerations here. If algorithms are to make decisions that rely upon open, contested and ambiguous values such as fairness, dignity or loss, then how are developers to define or embed these concepts? What is “fair” or “good” in technical terms?

The answer to these questions matter because the stakes can be very high—error or failure in ethically charged domains can have serious consequences for individuals or whole societies. A key governance question relates to the potential acceptance of algorithms that are known to be ethically imperfect but that are expected to improve through “learning by doing”. In what circumstances should failure lead to the outright rejection of particular techniques or advances, and in what circumstances should problems be managed with a view to arriving at better algorithms in the future? The answer may involve distinct governance measures such as public benchmarking, testing and certification.

2.3. Conclusion

The issues discussed in section 2.2 are at the heart of the evolving nexus of technology and governance in the machine learning field. However, as noted above, these overarching principles

operate differently, and warrant different responses, depending on domain-specific conditions. We can sketch out general areas of concern relating to machine learning, but the development of specific rules or regulations must be domain-specific.

In many instances, this is because machine learning algorithms are being deployed in domains with a rich pre-existing ecosystem of laws, norms, benchmarks, etc. The question here is to identify the relevant criteria, in terms of expected risks and benefits, and establish whether and how a decision-making algorithm should be evaluated against it. For example, in the criminal justice system, if “due process” is the foundational standard, then how is this to be embedded within machine-learning algorithms?

In the next chapter, we turn to consider three specific machine-learning domains: autonomous vehicles, public administration and healthcare.

3. Domain-specific applications and challenges

In the two preceding chapters we provided an overview of machine learning technologies and the overarching governance challenges they pose. We now turn to consider a number of specific domains where machine learning is being used. This is a crucial step—it is only at the domain-specific level that the analysis of general patterns and principles can be sharpened into the identification of specific challenges that require concrete governance responses. As we will see in the next chapter, there has been a flurry of national and international activity in recent years around the governance of artificial intelligence. A large amount of this activity has comprised overarching frameworks, guidelines and principles. In this chapter, we aim to demonstrate the need to go beyond such general principles to consider how they interact in concrete real-world contexts.

The three domains we consider are autonomous vehicles, public administration (with a focus on policing and criminal justice) and healthcare (where we focus on medical imaging, “software as a medical device” and precision medicine). In each of these domains we draw on chapter 1 to outline how machine learning is used in specific contexts, and we draw on chapter 2 to outline the key governance challenges that arise in each case.

3.1. Autonomous vehicles

The use of machine learning to enable autonomous vehicles promises a range of benefits, including improved safety, increased efficiency, more fluid traffic patterns, and a more personalised transport experience.

Machine learning in autonomous driving

The dominant approach to designing autonomous driving systems is to decompose the driving problem into smaller sub-problems and employ Deep Neural Networks (DNNs) to solve them individually. Broadly speaking, autonomous driving involves three types of tasks - perception tasks such as the detection and classification of lane-markings, traffic signals and pedestrians; planning tasks such as steering control, lane changing, stopping or slowing down; and mapping tasks such as localizing the vehicle to a map – all these to be undertaken in different weather conditions, visibility and road quality. Some of these tasks are performed by embedded ML-based systems, others rely on car-to-car or car-to-infrastructure communication. The former case is for when the car does not need active information sent to it by another car or by infrastructure, and can rely on information provided by its sensors only. In the latter case, the sytem relies on cloud-based solutions, and several constraints must be taken into consideration, including the speed of communication and privacy issues. Training a separate DNN for each task ensures diversity and redundancy in automated driving systems. The DNNs use the sensory data recorded during test

Illustration 6: A visual depiction of the salient objects learnt by DNNs
Source: (Bojarski, 2017)

Table 2: List of DNNs included in the NVIDIA DRIVE platform

DriveNet/ DepthNet	is used for obstacle and wait perception. It detects and classifies objects such as vehicles (including cars and trucks), pedestrians, traffic lights, traffic signs, and bicycles.
OpenRoadNet	detects free space around the vehicle. It distinguishes the boundary that separates obstacles from the drive-able collision-free space.
LaneNet	detects and classifies lanes, including the vehicle's ego-lane, adjacent lanes, and non-adjacent lanes.
MapNet	detects visual landmarks such as lanes and poles. It can detect features useful for path perception, as well as localization.
PathNet	predicts full geometry of drivable paths in images and on the three-dimensional road surface, regardless of the presence of lane markings.
WaitNet	detects an intersection and estimates the distance.
SignNet	classifies traffic signs detected by DriveNet, for US and EU.
ClearSightNet	determines if the camera view is blocked. It can predict three classes (clean, blur, blocked).
PilotNet	predicts driving center paths based on trajectories taken by human drivers.
Perceptive Automata	predicts human behavior by reading body language and other markers.
LightNet	classifies traffic lights (color, solid, and arrows) detected by DriveNet.

Source: (Nvidia, 2019)

drives from onboard sensors such as cameras, lidars, radars, IMUs and GPS to learn a representation that is best suited to their task. In some cases, the learning is self-supervised - for example, learning the control of steering-wheel using a dataset of road images paired with the steering angles generated by a human driver (Bojarski et al., 2016). Other tasks, such as lane identification and classification, require explicit labels that confirm the absence or presence of the different type of lane-markings in the training image (Wang, Ren, & Qiu, 2018).

The individual DNNs, though diverse in their learning approach, introduce redundancy due to the natural correlations that exist in the sensory dataset. For example, while training a DNN focussed on identifying pedestrians crossing the road, the learnt model will automatically include a representation of traffic lights - in images depicting pedestrians on the road the traffic light will almost always be present and will be red as opposed to green or orange. Thus, training different DNNs on the same sensory dataset leads to significant overlap in their learnings, thereby introducing much-desired redundancy when combined into a unified autonomous-driving system.

In machine learning parlance, this approach to combine individual models into a single optimised model is referred to as the Ensemble approach (Dietterich, 2000). In the case of autonomous-driving, an Ensemble approach increases the overall safety of the autonomous driving system due to the redundancy gains that the diversity of individual DNNs afford. For example, NVIDIA, a US-based company has developed an autonomous vehicle development platform (NVIDIA DRIVE) (Nvidia, 2019) that follows the ensemble approach to realize Level 2+ driving. Level 2+ adds surround perception and AI capability to SAE[11] defined Level 2 automated driving while maintaining the requirement that a human drives the vehicle. Illustration 6 shows the salient features learned by various DNNs employed in NVIDIA DRIVE. A list of DNNs used in NVIDIA DRIVE is also listed in Table 2.

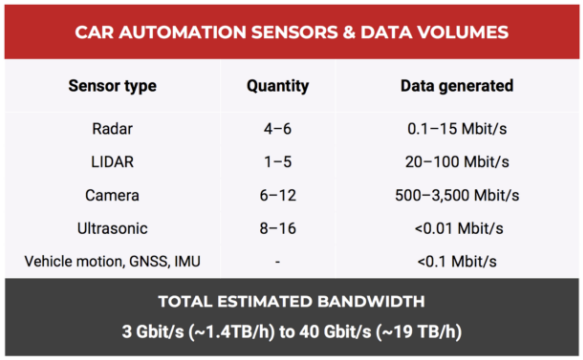


Illustration 7: An estimated 19TB of data generated every hour from an autonomous-driven vehicle
Source: (Rossi, 2019)

Table 3: A broad estimate of time taken to train a DNN on 10 million miles of data using a popular supercomputer

AlexNet	DNN for image classification tasks
ImageNet	Datatset of images for training tasks
DGX-1 (~120,000 USD)	Nvidia supercomputer with 8GPUs
Throughput	~150MB/s using single GPU
Convergence	50 epochs
10 million miles of training data	~6.3EB
Time to train	$\frac{6.3 \times 10^{18} \times 50}{8 \times 150 \times 10^{6} \times 3.15 \times 10^{7}}$ $= 82{,}687\ years$

Training DNNs for building an autonomous-driving system is no mean feat. In the real-world, traffic situations exhibit a long-tail distribution (Anguelov, 2019), i.e. a relatively small set of common situations, such as traffic lights turning green, yellow and red that occur quite too often, while a large set of diverse situations, such as stray animals on the street, construction material left unattended that occur only occasionally. To increase the likelihood of recording the infrequently occurring traffic situations, companies use fleets of test-driving cars that collect large amounts of data over extended periods of time. For example, Waymo, formerly the Google self-driving project, clocked 10 million miles of test-driving on public roads by August 2018 and 10 billion miles in simulation by July 2019 (Etherington, 2019). The array of sensors including radars, cameras, lidars, ultrasonic sensors, and other vehicle sensors that equip a vehicle body (Liu, Tang, Zhang, & Gaudiot, 2017) may generate up to 19TB of data every hour, see Illustration 7. Assuming Waymo test drives were conducted at an average of 30 miles per hour, this would suggest a

[11] https://saemobilus.sae.org/content/J3016_201806/

training data set of approximately 6.3EB (1 Exabyte = 10^{18} bytes). Further, following (Grzywaczewski, 2017) to assume the workload of training the DNNs for autonomous driving as similar to one of the computationally least intense image classification model - AlexNet (Krizhevsky et al., 2012) that achieves around 8*150MB/sec of throughput on a popular supercomputer - the Nvidia DGX-1, and achieves convergence in 50 epochs on a widely used dataset such as the ImageNet (Deng et al., 2009). Based on these assumptions, which are quite conservative, training an automotive DNN will take approximately 82,687 years on a supercomputer costing 120,000 USD, see Table 2 for calculation.

Companies like Waymo have invested in large-scale computing platforms that can train automotive DNNs in a much shorter time. These figures though approximate, provide an insight into the real challenges involved in training automotive DNNs. Further, there is no clear understanding regarding the number of miles of test-drive data required to provide clear statistical evidence of autonomous driving safety. As per the Bureau of Transportation Statistics in the US, on an average 1 fatality occurs for every 100 million miles driven by human drivers (Kalra & Paddock, 2016). Should an autonomous driving vehicle be test-driven for 100 million miles before it can qualify for commercial operations? Further, do the miles driven virtually in a simulated environment be included while evaluating safety criteria? Can regulatory incentives be designed to promote sharing of training datasets collected independently by different companies? Significant policy questions thus remain.

Meanwhile, Waymo became the first company to launch an autonomous commercial ride-hailing service, Waymo One, in the US in December 2018 (Krafcik, 2018). The service is currently limited to a select group of early riders in the suburbs of Phoenix, Arizona. Though designed to operate at SAE Level 4, i.e. no human driver attention is required, during this initial phase, Waymo One taxi come with human test-drivers to handle unforeseen situations.

Box 1 below provides some examples of policy initiatives regarding autonomous vehicles in the EU.

Box 1: Autonomous vehicles in the EU – examples of policy initiatives

The Federal Ministry of Education and Research (BMBF) in **Germany** has funded the Automated and Networked Driving project (EUR 100 million since 2015) to develop projects and initiatives for the exploitation of AI technologies for autonomous driving (Stix, 2018, p. 12). It is worth noting that Germany has a relevant market share for autonomous vehicles among EU members (Prescient & Strategic Intelligence, 2019), due also to the presence in the country of big players like Volkswagen or BMW Group and a huge representative of original equipment manufacturers (OEMs). Furthermore, Germany enforced specific legislation on those subjects starting from the Autonomous Vehicle Bill enacted in June 2017 (Autovista Group, 2019; Wacket, Escritt, & Davis, 2017), which defined the requirements “for highly and fully automated vehicles” such as the need for a black box recording the journey and logging whether the

human driver or the car's self-piloting system is in charge of the ride, thus, addressing the rights of the driver (Autovista Group, 2019; Wacket et al., 2017).

Other countries such as **France** and **Italy** enacted specific regulation, often close to "regulatory sandboxes" (Engels, Wentland, & Pfotenhauer, 2019). France allowed selected companies to test Level 4 vehicles (the total autonomy level being 5) on public roads around with no human driver at heavily restricted time and location (Autovista Group, 2019). Italy enacted a "Smart Road" decree on February 28 regulating testing of autonomous vehicles on 'specific roads' after authorization by the Ministry of Transport and Infrastructure (Autovista Group, 2019; Ministero delle Infrastrutture e dei Trasporti - Dalle info sul traffico all'assistenza alla guida, 2018).

In general, the EU legislation is paving the way to a fully autonomous car market (level 5 cars) that according to the Prescient & Strategic Intelligence consulting firm will be worth 191.6 billion by 2030, with a compound annual growth rate (CAGR) of 18.4% for the years 2023–2030 (Prescient & Strategic Intelligence, 2019).

Governance challenges

For autonomous vehicles, the primary success factor from the list discussed in section 2.2 is accuracy, in the sense of ensuring consistent performance standards that avoid accidents and thereby protect human safety. This means we are in a phase where technological improvements are the priority, for example in the area of real-time data quantity and quality (sensors, radar, cameras, LIDARs, etc.) and the overall quality of algorithmic decision-making. Robust testing and certification schemes are a further and related priority, in order to ensure the same level of consumer protection and confidence that apply for non-autonomous vehicles.

The development process for autonomous vehicles will entail an ongoing process of monitoring, feedback and adaptive improvements to increase accuracy. This process will inevitably involve errors and accidents; there have already been a small number of fatal accidents involving autonomous vehicles. Accommodating a certain level of vehicle error will be an unavoidable part of the development process. Although the clear priority is to arrive at a fail-safe scenario (i.e., designs that avoid failure), safe-fail scenarios (i.e., designs that protect safety even when failure occurs) are equally necessary during the development process because of the errors and accidents that are certain to occur. It is also important to bear in mind that errors are an unavoidable part of human driving. Full autonomy for a vehicle in real road conditions would require the vehicle to solve traffic problems that human drivers struggle with, such as another driver abruptly cutting in from an adjacent lane, or driving in snowy conditions that hamper braking and make road markings difficult or impossible to see.[12]

When autonomous-vehicle accidents occur, questions of algorithmic explainability and accountability (also discussed in section 2.2) will be an important element of determining legal

[12] For an example of how autonomous vehicles are tested on the road, see https://www.euroncap.com/en/vehicle-safety/safety-campaigns/2018-automated-driving-tests/

liability. At present, responsibility rests primarily with vehicle manufacturers when there is a software-related problem, but liability regimes are currently under review and will have to adapt as the technology develops and the relationship between vehicle and driver evolves. Accountability for errors in autonomous vehicles is therefore closely related to the criterion of human oversight (chapter 2), and whether the driver is determined to be in, on or off the decision-making loop at the time an accident occurs. The question of software liability and the responsibilities of software developers will need to be carefully analysed and reviewed as regulatory changes are considered. The insurance sector is likely to be an important player in the evolution of autonomous-vehicle governance and liability—insurers see many benefits from autonomous driving, and are working closely with car manufacturers and regulators to develop insurance solutions that will generate trust in the technology.

The Vienna (1968) and Geneva (1949) Conventions provide boundary conditions such as the requirement that a human driver can take control if needed. Remarkably, despite the fact that they were agreed before the age of autonomous driving, these boundary conditions allow automated and even autonomous cars on the road, provided that a human person is in control. Another remarkable aspect of these conventions is their continuous amendments over the years, as new technologies have come to the market. The core regulatory frameworks that apply to vehicle and road safety are being progressively updated to respond to the challenges posed by autonomous vehicles, particularly as they become increasingly embedded in the wider ecosystem of the Internet of Things (IoT). An important feature of the governance practices emerging in this area is the use of experimentation and “sandboxing”. There are also regulatory complementarities with other areas, outside the scope of this report, where AI/ML is being deployed, such as the manufacturing sector.

3.2. Public administration

This section discusses the use of machine learning in public administration. Public administrations are increasingly adopting new technologies in the management and provision of public services. Here we discuss the use of machine learning in the sensitive and ethically charged areas of predictive policing as well as aspects of the criminal justice system (such as whether a prisoner poses a risk of re-offending if released on parole).[13] These tools have the potential to reduce crimes and ease resource pressures in the criminal justice system. In Europe alone, as per a report by the Council for Penological Cooperation, the annual cost for penitentiary needs is 26 billion Euros (Aebi et al., 2016). Machine-learning applications capable of delivering even small improvements in crime prevention may lead to significant cost savings. However, they raise

[13] Other sensitive areas of public administration might include border controls and the allocation of public welfare benefits. In the latter area, a notable court decision was handed down in early 2020, when a machine-learning tool, SyRI (Systeem Risico Indicatie, or System Risk Indicator) was counter to European human rights and data protection rules. Four cities had been using the tool, which aggregated 17 categories of government data to identify individuals with a high risk of committing benefits fraud.

profound questions related to fairness, justice and the rule of law. In addition, their actual performance compared to a human benchmark is questionable.

Machine learning in predictive policing and criminal justice

PredPol is a US company, market leader in the development of a predictive policing tool that provides guidance on where and when to patrol. It uses supervised machine-learning algorithm that is trained on 2 to 5 years of crime data for each new city. The algorithm is updated every day

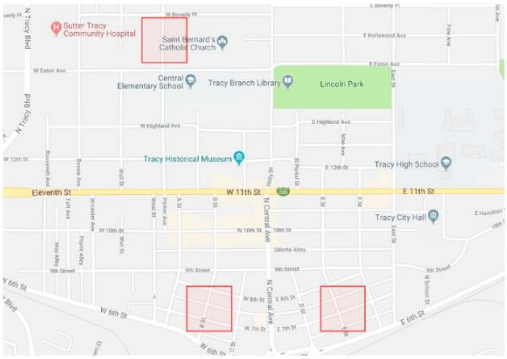


Illustration 8: PredPol algorithm predicting high-risk regions marked for priority-patrol for the day.
Source: predpol.com

with new events as they are received from the police department. It uses three data points – crime type, crime location and crime date & time. PredPol provides predictions by highlighting high-risk regions in a city in chunks of 150x150 meter square area. Around 50 police departments in the US and some in the UK use PredPol. Some police departments have claimed a reduction in street violence by 6% in a few months of adopting PredPol (Smith, 2018).

Though PredPol explicitly mentions that it does not use demographic, ethnicity or socio-economic information, there remain concerns regarding the generality of the criminological data used for training ML algorithms. Data is collected as a by-product of police activity, which is influenced by the high likelihood of some communities to call the police than others, and some crimes more likely to go unreported than others. As a result, ML algorithms tend to develop a bias towards certain types of crimes and regions, creating self-reinforcing feedback loops (Ensign et al., 2017). Some suggest reinforcement learning as an alternative approach for training ML algorithms for predictive policing (Haskins, 2019). Nevertheless, reinforcement learning requires a rule-based framework to evaluate each resource-allocation decision in an objective manner, thereby limiting the propagation of biases inherent in criminological data. Unfortunately, formulating such rules requires a holistic view of crime in a city – every single crime needs to be reported, and all types of crimes need to be pursued. While that is a big ask, for now, it is best to encourage private software providers such as PredPol to adopt greater transparency by making available the datasets they use for training their algorithms and allow the public at large to highlight anomalies and gaps in these datasets.

ML tools used for criminal risk assessment help make predictions, such when an individual could be released pending trial, or how likely a convicted criminal is to re-offend HART (Harm Assessment Risk Tool) is an ML-based tool used by the city of Durham, UK since 2013. Hart was initially trained on five years' worth of criminal data. It is used to predict if an individual is a low, medium or high risk. The city claims that HART's predictions were 98% accurate when predicting low-risk individuals, and 88% accurate when predicting high-risk individual. In the US, several cities use COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) an ML- based system that has till recently been reported as effective and satisfactory by the New York Division of Criminal Justice Services (DCJS, 2012). Nevertheless, a ProPublica investigation showed several instances of bias, especially against people of color (Angwin, Larson, Mattu & Kirchner, L., 2016). For example, Illustration 9 shows how two individuals, with very different prior offense history, when arrested for petty thefts received automated risk ratings that completely defy any conventional wisdom. A white male with serious crime history was awarded a significantly low-risk score as opposed to a black woman with history of juvenile misdemeanors was awarded a significantly high-risk score. After completing the sentence, the black women did not re-offend while the white male was subsequently booked for a grand theft.

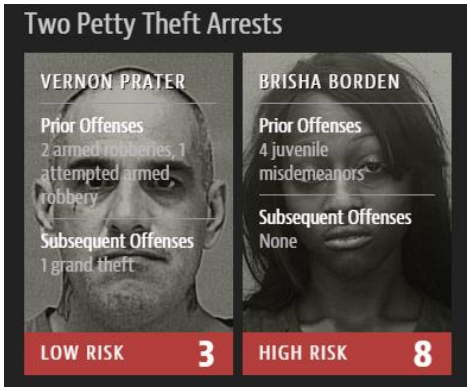
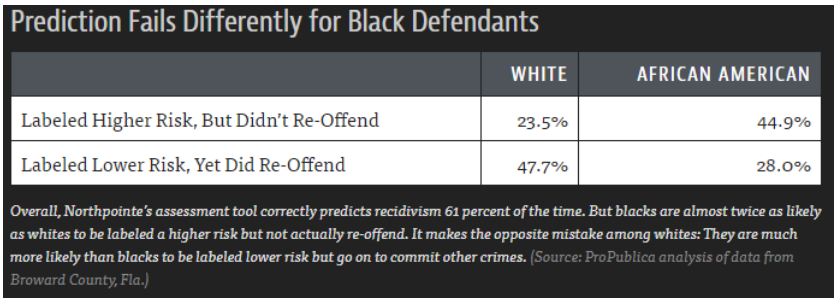


Illustration 9: Two petty theft arrests

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.
Source: (Angwin et al., 2016)

Over the years, several such investigations have revealed similar instances where subjective considerations have been discarded and individuals have been judged based on correlation with features identified in the training dataset. These correlations when interpreted as causations, as



Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Illustration 10: Prediction fails differently for black defendants
Analysis supporting claims of racial bias in a commercially available ML tool for criminal risk assessment
Source: (Angwin et al, 2016)

is the case with all ML-based prediction tools, lead to judgements based on stereotypes and not subjective considerations that individuals are entitled to. Illustration 10 summarizes ProPublica investigation that shows racial bias in commercially available ML tools for criminal risk assessment.

Box 2 below provides some exemples of policy initiatives regarding public administration governance in the EU.

Box 2: Public administration in the EU - examples of policy initiatives

The Commission's white paper *On Artificial Intelligence: A European Approach to Excellence and Trust*, identifies "parts of the sector" as being potentially high risk for the deployment of AI applications, mentioning in particular asylum, migration, border controls, judiciary, social security and employment services (European Commission, 2020a). More specifically, the white paper highlights particular concerns about the need to balance important interests in the area of facial recognition and similar remote-identification technologies. Using biometric data to identify a person is prohibited by the GDPR unless specific conditions apply, chief among them reasons of "substantial public interest" such as law enforcement.

However, the white paper stresses that such uses of these technologies are still subject to significant constraints: "It follows that, in accordance with the current EU data protection rules and the Charter of Fundamental Rights, AI can only be used for remote biometric identification purposes where such use is duly justified, proportionate and subject to adequate safeguards" (European Commission, 2020a, p.22). The Commission intends to launch a "broad European debate" to identify a common EU-wide approach on when these technologies can be used and on the safeguards to which they should be subject.

Governance challenges

In contrast to autonomous vehicles, the use of machine-learning systems in sensitive areas of public administration such as policing and the court system do not need to focus on technical factors of accuracy such as the ability to collect and process huge quantities data processing at

high speeds. Accuracy in a democracy's public administration is inextricably tied to fairness, reflecting the basic tenet of equal treatment that underpins the rule of law. Similarly, robust systems of transparency (de Laat, 2018), explainability, accountability (New & Castro, 2018) and human oversight are required—so that automated decisions that have legal effects on an individual can be understood, challenged and overturned where necessary, in the same way as non-automated decisions.[14] Finally, is is worth reminding that transparency and accountability are key aspects of administrative law. In the EU, this includes the "right of the institutions, bodies, offices and agencies of the EU [to] have the support of an open, efficient and independent European administration" created by Article 298(1) TFEU (European Commission, 2012), as well as Article 41 of the Charter of Fundamental Rights of the EU, which guarantees the "right to good administration" (European Union Agency for Fundamental Rights, n.d.). Consequently, the use of obscure ML technique in public administrations cannot be allowed whenever this may impinge on openness and accountability.

All of this reflects a structural difference between the use of machine-learning technologies in the public and private sectors. Public services typically apply uniformly across a territory to all the people who live there. Whereas a *consumer* can typically choose not to use a social media platform, for example, thereby avoiding being subject to whatever algorithms it uses, a *citizen* cannot typically choose whether or not to engage with core public administration functions. You cannot "delete your account" with the court system. This creates particular risks of lasting damage being caused, for example if algorithmic biases lead to incorrect predictions being made about an individual's propensity to offend. There are also broader concerns about the implications for freedom and autonomy. (Ferguson, 2018)

Automated decisions in the public sector need to comply with the fundamental rights and values embodied in a country's constitution, laws, treaties, and so on.[15] Values such as privacy, non-discrimination and human dignity are cited in the EU context, for example. (European Commission, 2020a, p.17) Of course, there will be potentially significant differences between countries and regions depending on the values that prevail. In broad terms, the EU is more wary than the US to use algorithmic decision-making for public administration, while China is more willing than both the EU and US to do so. The strong ethical focus of the EU in this area gives it a potential advantage in terms of setting global norms for the responsible use of machine learning in sensitive areas of public administration. There is an analogy here with data protection and the global influence that the values embodied in the GDPR have had. In chapters 5 to 7 we discuss this at greater length, asking also whether EU influence on the global governance of machine

[14] In the European Commission's 2020 White Paper on AI, the definition of high-risk applications refers to "AI applications that produce legal or similarly significant effects for the rights of an individual or a company" (European Commissiona, 2020).
[15] For a discussion of how public bodies can "tap into new sources of data and conduct statistical inference and causal inference in a principled manner" (Anastasopoulos & Whitford, 2018).

learning in ethically sensitive domains might also create commercial advantages for European developers of machine-learning solutions in these areas.

3.3. Healthcare

Health care is a complex sector in which many applications of machine learning are considered with great expectations. It is used or considered in various sectors or applications, including patient intake and referral (self-service initial consultation), hospital triage and prevention, medical image interpretation, software as a medical device, and precision medicine (genetic diagnostic and prediction, therapies). Its potential for future applications is tremendous. Machine learning applications, can *help* interpret results and *suggest* diagnoses, and predict risk factors to *help* introduce preventative measures. They can also *suggest* treatments and *help* doctors create highly individualised treatment plans. Combined with the knowledge of doctors and other medical experts, AI can lead to better accuracy, higher efficiency and more positive outcomes in the health field.

At the core of these advancements in medicine, and fuelling the development of AI, is health or health-related data, i.e. data collected from various sources that may describe or influence a person's state of health. Health data includes information on people's health (e.g. symptoms, allergies, impaired vision or hearing), on health care (e.g. medications, surgical procedures), data from conventional investigations (e.g. blood pressure, laboratory values, electrocardiography, X-rays), and results from genetic or other laboratory tests. Health data also involves information on lifestyle (e.g. nutrition, alcohol and drug use, smoking, exercise) and data on the environment (e.g. air and water quality, second-hand smoke, exposure to noxious substances) as well as insurance data and socio-economic data.

This section focuses on three applications of machine learning in the healthcare domain: computer-aided imaging interpretation, software as a medical device, and precision medicine. In each of these three areas case, we discuss how machine learning is used and highlight the specific governance challenges that arise. The section closes with a discussion of the overarching governance challenges for machine learning in the healthcare sector. It is one of the most tightly regulated sectors—particularly in terms of safety, efficacy and cost-effectiveness—and regulations are already being adapted in response to advances in diagnostics and treatments.

Medical imaging

In many health care scenarios, imaging has become the default diagnostic option due to the greater availability of imagining devices and advances in capabilities. As a result, the quality of care is dependent on the capacity to interpret imaging data quickly, with low error rates, and with a wide diagnostic range. Highly capable ML-based algorithms have been developed for medical image interpretation, notably to identify breast (and other) cancer tissue (Gupte, 2019; Rodriguez-Ruiz et al., 2019). ML-based image interpretation is one of the first applications of ML (Towers-

Clark, 2019). Adoption is still low but is anticipated to grow fast, for a wide range of diagnostics, in parallel to increase in accuracy, reliability and overall performance of the ML algorithms that support them.

There are various imaging methods available for the screening and diagnosis of breast cancer, including mammography, ultrasound, thermography and more recently tomosynthesis. Nevertheless, image analysis may be adversely affected due to the presence of noise in images, poor contrast, inadequate clarity and the radiologist's lack of experience or restricted visual-perception ability. For example, the presence of clusters of tiny deposits of calcium, called microcalcifications (MCs), in mammograms may indicate early-stage breast cancer. Individual MCs are often difficult to detect due to their relatively small size (0.05-1mm), variation in shape, orientation, brightness and the confounding texture of surrounding breast tissues. Machine learning techniques such as Support Vector Machines (SVMs) have been quite effective at detecting MCs in a mammogram, see Illustration 11, and have been widely employed as a second reader alongside a certified radiologist (Rao et al., 2010). The SVMs are trained in a supervised manner using a dataset composed of hand annotated mammograms where radiologists highlight regions in the mammograph with "MC present" and "MC absent" tags.

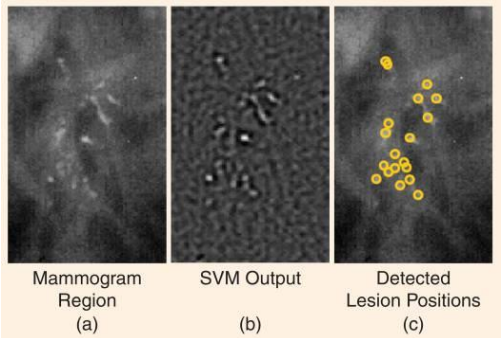


Illustration 11: An example mammogram.

It containing microcalcifications (a), output of SVM based detector (b), detected MC positions obtained by thresholding (c).
Source: (Wernick, 2010)

While mammograms are used to detect cancers, there is also a risk of false positives that can lead to unnecessary biopsies and surgeries. In addition to assisting radiologists in reading mammographs, ML is also being used to predict if a breast lesion is benign (non-cancerous) and can be safely surveilled, or it is malignant (cancerous) and requires surgery (Massat, 2018; Kooi, et al., 2016) used a Convolutional Neural Network (CNN), a deep learning technique specialized for image analysis, to train on a set of around 45,000 images and achieve similar performance as a group of three experts including experienced readers and certified radiologists. The mammograms used were collected from a large-scale screening program in The Netherlands (RIVM, 2019). All tumours in the dataset are biopsy-proven malignancies and annotated by an experienced reader. ScreenPoint is a Dutch medical imaging company that is at the forefront of commercializing ML-based mammography reading software. ScreenPoint's software, Transpara

(ScreenPoint Medical, 2019), uses CNNs trained with over 1 million mammograms. Other ML-based techniques, such as the swarm-optimized wavelet neural network, PSOWNN (Dheeba, 2011) have also been considered. PSOWNN reported 92% accuracy in identifying breast cancer using mammograms.

Other examples of use of ML in medical imaging exist, including the diagnosis of Alzheimer using PET scans (Ding, 2018) and pneumonia from chest x-rays (Rajpurkar et al., 2018).

As per (WSJ, 2017), around 800 million radiology exams are conducted annually in the US, generating almost 60 billion images. Europe too relies heavily on the use of radiology for diagnostics. It comes as no surprise that hospitals struggle to effectively store, display and distribute these images throughout the enterprise. A typical imaging workflow is composed of interactions between the following three IT systems: Radiology Information System (RIS) – for tracking radiology orders, Picture Archiving and Communication System (PACS) – for storing and digital transmission of electronic images, and Clinical Information System (ICS) – for aggregating information doctors need for patient care. To ensure seamless integration, the interactions between these IT systems follow the Digital Imaging and Communications in Medicine (DICOM) standard (MITA-NEMA, 2019). DICOM is an international standard that describes the means of formatting and exchanging medical images and image-related information to facilitate the connectivity of medical devices and systems (Siemens-Healthineers, 2019). As the adoption of ML in analyzing medical images grows, standards like DICOM will have to be extended to standardize the exchange of information between machines running ML models and the hospital IT systems. The concerns are similar to the ones listed in the concluding section of Chapter 1. Seamless integration of ML systems with production-grade medical imaging systems will need to address audit and compliance requirements, while maintaining the privacy and security of health records.

Specific governance challenges and implications

The key governance priority in this area is human oversight. As things stand, the software produces an analysis that is used as a decision-support—decisions are not (and do not need to be) automated. Doctors are in the loop. As long as this remains the case, these algorithms are relatively low-risk and uncontroversial. However, that may change as the technology evolves—if, for example, machines are able to identify clinically salient patterns in images at a level of detail that are no longer human-readable. The removal of the doctor from the loop would have important implications, particularly in terms of explainability and accountability.

Another area where human involvement is an important parameter is in the training of machine-learning models. Most machine-learning applications in medical imaging involve supervised learning, but the nature of the human role in the training process can have significant implications:

- Assigning labels to confirm the absence or presence of specific features in a medical image is potentially prone to human error and bias, which could then be learned by the algorithm
- If the labelling process focuses not on the features of each training image, but on the disease-related outcome associated with each image, this requires little or no human judgement. It also potentially allows the machine to identify new patterns in the images that go beyond existing human knowledge, allowing for new predictions/diagnoses. However, the price to be paid would be a relatively opaque model, requiring additional support to meet interpretability and accountability requirements.

Another crucial governance challenge in this area is privacy. The success of machine-leaning algorithms rests on the availability of large training datasets. In the medical imaging domain this has led to various alliances and acquisitions involving technology companies and medical services providers. For example, in 2015 IBM acquired Merge Healthcare to gain access to some 30 billion medical images from 7,500 hospitals and clinics in the US (MIT Technology Review, 2015). From a governance perspective, this raises serious questions about the ownership of medical records. For example, does a radiography image belong to the patient, doctor, equipment manufacturer or hospital? Can it be traded, even when anonymized, without explicit consent?

Software as a medical device

Medical devices increasingly rely on a piece of sophisticated software. If the device has the ability to interact with the patient's health (i.e., if it does more than collect information about the patient and send it elsewhere, which is the case with pacemakers and insulin pumps for example) it is regulated. This means it requires specific authorisation before being placed on the market. In both the US (FDA, 2018) and Europe, the regulation of “software as a medical device” (SaMD) is currently under review. The regulatory challenges are intensified when the software involves machine learning.

Machine learning techniques coupled with advancements in IoT (Internet of Things) design have enabled rich and connected information ecosystems that can be monitored and controlled with great precision and in real time. In the healthcare sector, more and more medical devices are being designed as “closed-loop” systems that interact directly with the patient, such as automated insulin delivery systems (often referred to as the artificial pancreas). In June 2017, Medtronic, a medical device company, introduced an automated insulin pump system—MiniMed670G—the first technology approved by the FDA for the automated delivery of insulin to diabetes patients. It uses a proprietary algorithm to automatically adjust the dosage of insulin every five minutes based on continuous measurement of the patient’s glucose level using a small sensor-device worn just under the skin. The MiniMed670G is a “hybrid closed-loop” system. One reason for this is that the sensor needs periodic calibration using a standard blood glucose monitoring device. However, Medtronic has launched trials for MiniMed780G with the aim of providing a truly closed-loop

automated insulin system that will require no calibration and that will deliver insulin based on predictive analysis developed jointly with IBM Watson Health. (At present this predictive analysis is available through an app, Sugar.IQ, that uses data from the sensor, meal logs and other user inputs).

Another example is Philips Healthcare's IntelliVue Guardian Solution, a patient monitoring system that uses AI to predict when a life-threatening crisis may occur in a patient for effective and early intervention. It includes an Early Warning System (EWS) that combines software, clinical decision support algorithms and mobile connectivity to harness the data received from wearable devices. For example, a clinician can place a cableless device embedded with sensors on a patient's wrist to track vital signs such as blood pressure. The IntelliVue Guardian Solution software uses machine learning to identify any significant changes in the patient's vital signs based on algorithms trained on large datasets of similar patient data. If an important change is identified, data would be transmitted to IntelliVue monitors or mobile devices to notify the caregivers.

Specific governance challenges and implications

As with autonomous vehicles, accuracy is a key governance criterion for SaMD, in terms of preventing errors that might adversely affect patient health. Machine learning-based SaMD presents a particular challenge for regulators, because of the learning process. The software is designed to evolve in response to real-world data, but this evolution might mean that after a series of iterations the software has changed significantly relative to the version that was initially approved by the regulator. The regulator's core responsibility is to ensure that the software will remain safe, but there are also strong reasons to permit modifications that will increase the performance and safety of the SaMD, without necessarily requiring every iteration to receive a new authorisation.

In the US, the FDA is considering a new regulatory framework that would apply to "modifications to AI/ML-based software as a medical device". (FDA, 2020) At the heart of the proposed framework is the idea of a "pre-determined change control plan", which would set out two things in advance:

- the types of modification that are envisaged
- a protocol governing how these modifications will be implemented in a way that manages risks to patients

European regulators will need to consider similar regulatory initiatives in order to maintain and extend their influence in this area of healthcare-technology governance.

Precision medicine

Precision medicine is an approach to medicine that moves away from the "one-size-fits-all" approach to healthcare delivery and tailors treatment and prevention strategies to people's unique

characteristics, including environment, lifestyle and biology. It allows for detailed molecular characterisation of many disorders and diseases via sequencing of patients' DNA.

Machine learning algorithms can accurately analyse the sequenced information and leverage the gigantic amount of data in an individual's medical records with direct benefit for the patient. This helps physicians to make better decisions and create more effective treatment plans. Machine learning is thus absolutely central to improving patients' health and medical treatment with precision medicine ("there is no precision medicine without AI"), but still in its early stages (Insights Team, 2019; Papadakis et al., 2019).

Genomic data holds immense potential to enhance health care for patients suffering from hereditary diseases and cancer. Six percent of the global population is affected by a hereditary disease and it is expected that one in two people will be diagnosed with cancer in the coming years (CRUK, 2015). In both cases, clinical genomic analysis is useful to pinpoint the molecular origin driving the disease and to tackle the cause of such disease rather than act on its consequences. Researchers are using machine learning to identify patterns within high volume genetic data sets to help identify genetic variations affecting crucial cellular processes, including metabolism, DNA repair, and cell growth. Disruption to the normal functioning of these pathways can potentially cause diseases such as cancer.

One example is the use of ML-based genomic analysis for lung cancer treatment. In the past treatment was prescribed based upon the patient's tissue type instead of a particular genetic mutation. Using AI-based solutions such as the one from SOPHiA Genetics (SOPHiA, 2019), a Swiss company that specializes in clinical genomics, physicians can now identify the genetic events that caused the condition in the first place, and not just treat symptoms. The process begins with the extraction of the patient's DNA via a blood draw or biopsy. The hospital then uses molecular biology processes to prepare the samples and subsequently digitize them using a DNA sequencer. The resulting genomic data is then submitted to AI systems on the Sophia DDM Software as a Service (SaaS) platform, which uses statistical inference, pattern recognition and machine learning techniques to identify the patient's genomic mutations. The more hospitals use the analytics platform, the more patients' genomic profiles are accumulated, and the smarter the platform gets. Without this technology, the process of determining a drug treatment takes about two days of work, and in some cases can take several months when using old technologies. In contrast, healthcare professionals who use Sophia for genetic sequencing and analysis can get drug treatment regime recommendations for individual patients in one day. One benefit of the Sophia system being a SaaS-based solution is that even smaller hospitals and clinics can afford the technology, which on average costs $50-$200 per genetic evaluation. To ensure patient-privacy, references to individual patients are stripped-off of all treatment records so that the data is fully anonymized before it leaves the hospital IT system. SOPHiA's technology for clinical

genomics is being used by geneticists, pathologists and radiologists in over 990 leading hospitals worldwide.

Another example is the use of machine learning in determining the appropriate dose for immunosuppressive agents, such as Tacrolimus that is widely used to prevent acute rejection following a renal transplant. Tacrolimus must be administered with utmost care. Insufficient dosing of tacrolimus can lead to acute rejection, while overexposure can lead to drug-related toxicities, such as nephrotoxicity, neurotoxicity, and new-onset diabetes. The pharmacokinetics of tacrolimus depends on several factors including clinical genetic factors such as CYP3A5, CYP3A4 and ABCB1 single nucleotide polymorphisms (SNPs). Machine learning has been found to be particularly useful over conventional statistical techniques due to its ability to model non-linear effects and interpretation of large genomic data sets. (Tang et al., 2017) provides a comparative study on the suitability of various machine learning algorithms to tacrolimus dose prediction.

Application of ML to precision medicine though very encouraging, has a long way to go before being widely adopted in clinical use. The complicated situation of real patients should be well considered, as additional comorbidity and the effect of interacting drugs may not be completely included in the models. For example, in the tacrolimus dosage prediction study, additional factors such as smoking, alcohol consumption and data on adverse reactions were not included in the training process.

Direct-to-consumer use of genomics services is also growing rapidly. A good example is the Genetic Weight Report by the personal genome services provider 23andMe (23andMe, 2017). Body weight is determined by a complex combination of lifestyle, environment, and genes. Genetic Weight report uses machine learning techniques to calculate a person's genetic predisposition to weigh more or less than average. The report also provides insight into which lifestyle factors might make the biggest difference for a person's own weight. The model is trained on a collection of 1 billion phenotypic and genotypic data points from its 10 million customers. 23andMe uses this massive data to conduct research that could lead to new insights and treatments into conditions like cancer, dementia and diabetes. In 2017, 23andMe was granted first ever FDA authorization to market direct-to-consumer Genetic Health Risk reports, including tests for Alzheimer's and Parkinson's disease.

Box 3 below provides some exemples of policy initiatives regarding the healthcare governance in the EU.

Box 3: Healthcare in the EU - examples of policy initiatives

In the healthcare sector the European Commission has stressed that a successful digital transformation requires "appropriate regulatory frameworks that will safeguard the rights of the individual and society" and "full compliance with data protection legislation and ethical principles" (European Commission,

2018b). The Commission's white paper on AI (European Commissison, 2020) identifies healthcare as a potentially high-risk sector for the application of AI. Finland has sought to flesh out in more detail how AI should be used in the healthcare sector, and has listed ten principles in a code of conduct for the use of AI in healthcare:

1) Understand users, their needs and the context.
2) Define the outcome and how the technology will contribute to it.
3) Use data that is in line with appropriate guidelines for the purpose for which it is being used.
4) Be fair, transparent and accountable about what data is being used.
5) Make use of open standards.
6) Be transparent about the limitations of the data used and algorithms deployed.
7) Show what type of algorithm is being developed or deployed, the ethical examination of how the data is used, how its performance will be validated and how it will be integrated into health and care provision.
8) Generate evidence of effectiveness for the intended use and value for money.
9) Make security integral to the design.
10) Define the commercial strategy

However, although there is evidence of Europe setting out clear ethical priorities in the healthcare sector, it is worth repeating a point from chapter 3 about the opening up of a potential regulatory gap between the US and Europe in respect of the most advanced (and ethically challenging) healthcare applications of machine learning. The FDA is currently considering the regulatory implications of AI/ML-based software-as-a-medical-device (SaMD), as machine learning is used in such a way that the software/device is self-evolving. (FDA, 2020) If Europe aims to be a leading influence in the global governance of machine learning in the healthcare sector, then it would need to adopt a similarly horizon-scanning approach towards the need for regulatory preparedness for new/emerging technologies and applications. The EU is in the process of a major regulatory transition in the area of medical devices, with the news Medical Devices Regulation (MDR) (Regulation (EU) 2017/745, 2017) set to take effect in 2020, but it is not yet clear how this new framework will deal with cases such as self-evolving algorithmic devices.

Specific governance challenges and implications

Prevision medicine raises a number of the governance challenges we discussed in chapter 2. Privacy concerns are particularly prominent, because personalized advice or treatment requires access to all of a person's relevant data, including sensor, genome, microbiome and medical records. (One practical obstacle to the collection of such comprehensive datasets is that in many countries people don't have ready access to their medical records, while most health-related data is spread across multiple systems owned and managed by device manufacturers and service providers.) Another data-related challenge concerns the training process: For a machine learning system to be validated and refined in clinical studies it is crucial that it is trained on diverse datasets that are representative of the target population(s).

As with all healthcare applications of machine learning, accuracy is a paramount concern. One of the benefits of precision medicine is that the accuracy of the diagnostic and treatment can be calibrated at the level of the individual patient. However, achieving this kind of targeted accuracy

involves a potential trade-off with explainability—it is impossible for a human to grasp all of the details that lead to a precision-medicine diagnostic. This in turn has implications for human oversight. On the one hand, a human doctor is still firmly in the loop when a precision-medicine diagnostic is used in the clinical decision-making process. But on the other hand, the diagnostic may be opaque to the doctor because of its machine-learning origins.

This growing pressure on doctors to deal with rapidly expanding "responsibilities of data management and analysis" is one of a number of challenges to the adoption of machine learning in precision medicine that are highlighted by Xu et al (2019). Others include: difficulties in establishing a baseline against which to validate the real-world usefulness of machine-learning systems; and problems of transparency and reproducibility related to the lack of detailed information about the specific machine-learning techniques and models used in a product or service.

Finally, it is worth noting that the rapid growth of the direct-to-consumer genomics industry raises specific governance challenges. There are serious concerns about regulation in this area and about the role of health professionals in helping individuals to interpret their test results (Goldsmith et al., 2013). One recommendation is that health professionals be provided with training to ensure that they are familiar with the output of consumer genetic tests, so that they can support patients who ask for help interpreting and utilizing the test results.

4. The current governance landscape

This chapter surveys the landscape of machine-learning governance. Against a backdrop of proliferating activity in this area—particular in the broad area of non-binding frameworks for "ethical AI"—we review a range of documents that have been produced by key governments and other organizations. Our aim is to analyse the predominant perspectives, priorities and recommendations in this area, with a view to highlighting the key characteristics that distinguish the governance of machine learning in the European Union from that in other countries and regions. We begin with a discussion of governance strategies in three jurisdictions that are driving global developments—the EU, US and China—before turning to consider other national strategies, intergovernmental initiatives and select developments in the academic, scientific, private and non-profit sectors.

4.1. The three major players

The European Union

While the EU has been actively considering the implications of AI-relevant technological developments for more than a decade (Renda, 2019), the flow of initiatives focused specifically on AI has picked up pace since April 2018, when 25 European countries (24 of the EU's member states, and Norway) signed a "Declaration of Cooperation" on AI which highlighted three broad priorities: (i) building technological and industrial capacity, (ii) ensuring labour-market and educational inclusion, and (iii) establishing a legal and ethical framework that builds on fundamental EU rights and values (European Commission, 2018d). This political agreement among states was followed within weeks by a European Commission Communication on "Artificial Intelligence for Europe" (European Commission, 2018a). This document placed the EU's approach to AI firmly within the context of international strategic/geopolitical competition:

> *"Like the steam engine or electricity in the past, AI is transforming our world, our society and our industry. Growth in computing power, availability of data and progress in algorithms have turned AI into one of the most strategic technologies of the 21st century. The stakes could not be higher. The way we approach AI will define the world we live in. Amid fierce global competition, a solid European framework is needed." (ibid., p. 2)*

The three-fold objectives of the "Declaration of Cooperation" are rehearsed in the Commission's Communication, but there is a notable focus on the third, values-based component. For example: "The EU can lead the way in developing and using AI for good and for all, building on its values and its strengths" (ibid, p. 3). Or elsewhere: "we can place the power of AI at the service of human progress" (ibid., p. 20). This idea of a distinctively European approach to AI is echoed in another 2018 document, "Artificial Intelligence: A European Perspective", published by the Commission's

Joint Research Council, which again stressed the backdrop of international competition before discussing the AI-related challenges facing the EU across eight dimensions: ethical, legal, educational, economic, cybersecurity, energy, data, and societal resilience (Annoni et al., 2018).

The normative focus of the EU's approach to AI was further emphasized in 2019 with the publication of a set of "Ethical Guidelines for Trustworthy AI" by a High-Level Expert Group on Artificial Intelligence (AI HLEG) that the Commission had established (AI HLEG, 2019; Floridi, 2019). These guidelines rest on the following "ethical imperatives": (i) respect for human autonomy, (ii) prevention of harm, (iii) fairness, and (iv) explicability (AI HLEG, 2019, p. 12). And the group provides a (non-exhaustive) list of seven requirements for trustworthy AI (AI HLEG,2019, p. 14):

- human agency and oversight
- technical robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- societal and environmental well-being
- accountability

With these initiatives, the EU aims to develop its own approach to dealing with the potential tensions created by machine learning and its application to automated decision-making between, on the one hand, innovation and growth, and, on the other hand, concerns for ethics, autonomy and well-being of human individuals and communities. As this chapter will highlight, there are some important differences between views and priorities of organisations of the European Union and those from non-European governments or some major players. It is worth noting that the EU strategy overall pays specific attention to the development of AI and ML in government, as an important component of public sector innovation.

As things stand, there is no EU regulation of AI or machine learning per se, but moves in that direction are being made. In February 2019 the European Commission launched a public consultation with a white paper (European Commission, 2020a) setting out, among other things, "the key elements of a future comprehensive European legislative framework for AI in Europe". The white paper adopts a risk-based approach to regulation in this area and notes that the idea of risk should be understood here with "in particular from the viewpoint of protection of safety, consumer rights and fundamental rights". It goes on to specify two cumulative criteria for identifying most high-risk applications of AI:

- the AI application is used in a sector where significant risks can be expected to occur
- in addition, the AI application is used in such a way that significant risks are likely to occur

If these two criteria apply—or in other "exceptional instances" that justify a high-risk designation—then an AI application would be subject to a series of mandatory requirements, which the European Commission suggests would focus on the following areas:

- ensuring that the training data used to develop AI systems conform with the EU's values and rules
- keeping all records and data needed to allow potentially problematic decisions by AI systems to be traced back and verified
- proactive provision of information about high-risk AI applications to authorities and affected parties
- all reasonable measures should be taken to ensure robustness and accuracy over the all life cycle phases of the AI system
- all high-risk AI applications would require appropriate human oversight
- special requirements would apply to the use of biometric data for identification process—such as facial recognition—because of specific risks such AI applications pose to fundamental rights.

Unless and until the White Paper on AI results in binding regulatory action, the EU-wide General Data Protection Regulation (GDPR - European Commission, 2016) remains a keystone of the EU's response to the risks and challenges of increasing digitalization in most domains of life. The GDPR took effect in 2018, but it builds on decades of EU regulatory action in the area of data protection, which is enshrined as a fundamental right in the EU treaties.

Although it focuses primarily on data protection issues—setting out principles governing the collection and processing of personal information—the GDPR also has direct implications for machine learning. For example, it makes some provision for (i) an individual right not to be subject to automated decision-making, and (ii) a right to information and/or explanation when such automated decision-making occurs. On the first of these rights, Article 22.1 states: "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her". However, there is some ambiguity as to when this right will be applied. The GDPR lists a number of factors that can overrule the right, including how it is interpreted by individual EU member states. According to Article 22.2, the right does not apply if the automated decision in question: "a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; b) is authorized by Union or Member State law…; c) is based on the data subject's explicit content."

In the Recitals (non-binding clarifications) to the GDPR, the idea is introduced of an individual right to explanations of how algorithmic decisions have been reached. Recital 71 of the GDPR notes that Article 22.1 should not be interpreted as prohibiting algorithmic decision-making, but goes on to state that: "[automated] processing should be subject to suitable safeguards, which

should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision processing." (ibid.) A similar point is made in Article 13 and elsewhere, stipulating that individuals must be provided with information concerning "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject". However, there is no consensus among scholars as to the extent to which the "right to an explanation" can be realized. See for example Wachter et al (2017), Doshi-Velez and Kim (2017), Malgieri & Comandé (2017).[16]

It is worth noting that a number of other pieces of EU legislation with relevance to machine learning are currently under review. These include the 2006 Machinery Directive, which is being assessed to ensure that it is fit for purpose given the increasing role of machine-learning algorithms in machines of all kinds. In addition, in March 2018 the European Commission established an expert group on "liability and new technologies", one sub-group of which is developing guidelines on the application of the 1985 Product Liability Directive to AI/ML.

The US and China

While there are areas of overlap in the approaches to AI being taken by the EU, China and the US, there are sharp differences of emphasis. Broadly speaking, the US and China lean towards prioritizing innovation and growth while the EU leans towards values-based protections (for a more detailed discussion, see Renda, 2019, p. 39).

The United States

The US government takes a "hands-off" approach towards the development of AI in general as well as in the specific field of machine learning technology (ibid., p. 38). This can be explained by the fact that the country is the global leader in the field, enabling its technology sector and universities to engage in extensive research and development activity without direct government support or encouragement. However, as international competition has intensified, and as the dividing line between commercial and geopolitical rivalry has become more porous, there are signs of the emergence of a more "hands-on" approach to the development of AI. As for these issues, in February 2019, the US President issued an executive order called the "American AI Initiative", as a coordinated federal government strategy (White House, 2019). Staying at the front of the global AI race is at the heart of this initiative: "It is the policy of the United States Government

[16] See also Poullet (2018, p. 778) on the potential significance of the exclusion of anonymous data from the terms of the GDPR.

to sustain and enhance the scientific, technological, and economic leadership position of the United States in AI R&D and deployment." The initiative comprises five principles:

- drive technological breakthroughs
- ease the creation of new industries and the adoption of AI by existing industries
- prepare for skills and labour-market challenges
- foster trust in AI and protect civil liberties, privacy and American values
- promote an international environment that opens markets for US AI industries

While there is clearly an ethical component to this US strategy, it is much less pronounced than in the EU initiatives discussed in the previous section. This reflects familiar differences in political and commercial culture between the EU and the US. Comparison can be made to other policy areas in which the EU has prioritized a "precautionary principle" that requires protective measures to be taken where there is scientific uncertainty as to the risk of severe, potentially irreversible damage. By contrast the US tends to allow innovation to proceed unless there is clear scientific evidence of harm.

However, the extent of any divergence between the approaches towards AI/ML that are being taken in the US and Europe should not be overstated. In the US there are indications across politics, industry, science and non-governmental organisations (NGOs) that ethical concerns about the consequences of AI applications are rising up the agenda. For example, several high-profile Congressional hearings about private data access, sharing and use by social media platforms have served to draw attention to the issue of privacy. In April 2019 a group of Democratic Party lawmakers tabled a draft bill on "algorithmic accountability", which would require large firms that process large volumes of personal information to evaluate their machine-learning systems for adverse impacts on "accuracy, fairness, bias, discrimination, privacy and security" (The Senate of the United States, 2019).

China

One of the factors adding momentum to the evolution of a more hands-on approach to AI in the US has been the pace of developments in China, where the government has set itself an unambiguous three-step objective for achieving global AI pre-eminence. As set out in the "Next Generation Artificial Intelligence Plan" published in 2017, the objective is to catch up with other advanced economies by 2020, to make key AI breakthroughs by 2025, and to become the global AI leader by 2030 (Dutton, Barron & Boskovic, 2018).

The centralized structure of China's political economy and society mean that —in contrast to the US or the EU—it is relatively straightforward for the authorities to ensure that an overarching goal is cascaded throughout lower-levels of government, industry and society. China's AI plan comprises a range of coordinated advances in areas including: education, finance, industry, the

military and the judiciary (Fischer, 2018). A supplementary detailed action plan sets out four key tasks for the period 2018-2020:

- intelligent and networked "smart" products in eight categories, including vehicles and identification systems
- advanced technological underpinnings for AI, such as sensors and neural network chips
- "intelligent manufacturing"
- improve the overall AI ecosystem with measures such as training, standards and increased cybersecurity (Dutton et al., 2018)

The ethical backdrop against which China's AI strategy is being rolled out is starkly different from that in the EU. Chinese society focuses on the community rather than the individual—at an OECD discussion of the social and economic implications of AI, "the individual-centred cognitive approach" to AI/ML in western societies was contrasted with "the society-centred connectionism" adopted in Asia (OECD, 2017b). While, as we have seen, the EU has put protection of individuals' data at the heart of its evolving digital policy, one of China's means of accelerating progress towards its goal of AI leadership is to ensure "massive data availability for machine training", even if this requires the use of "very intrusive technological means" which may "end up sacrificing the protection of personal data on the altar of faster, more capable machines" (Renda, 2019, p. 39).

4.2. Other national initiatives

We now turn from the global interplay of AI/ML strategic and policy dynamics in the EU, the US, and China to consider a range of national AI/ML plans that have been announced by governments around the world. The results of this analysis are summarized in Table 4 below, which highlights the emphasis that different countries are placing on six recurring dimensions (adapted from Dutton et al., 2018). These are: scientific research and talent development, industrialization, ethical standards and regulations, data and digital infrastructure, AI in government, and inclusion and well-being.

Table 4: Strength and focus of the strategies for AI and machine learning in selected countries

Countries	Dimensions					
	Scientific Research and talent development	Industrialization	Ethical standards and regulations	Data and Digital Infrastructure	AI in Government	Inclusion and Well-Being
Europe / Eurasia						
EU						
Denmark	x	x			x	
Estonia				x	x	
Finland	x	x				
France	x	x	x	x		x
Germany	x	x	x	x		x

Italy	x		x	x	x	
Sweden	x					
Switzerland	x					
UK	x	x	x	x	x	x
Russia	x	x				x
Middle East						
Israel	x	x				
Saudi Arabia		x		x		
Emirates (UAE)		x		x		
North America						
USA			x			
Canada	x	x	x	x	x	
Asia						
China				x		
India			x	x		
Japan		x	x	x		
South Korea	x	x	x	x	x	x
Oceania						
Australia	x	x	x	x	x	x

Note: The colour-coding of the cells works as follows: light shading indicates that the dimension is covered in a country's strategy; dark shading indicates a strong emphasis on the dimension; an x indicates a special focus on machine learning specifically. (Note that for some countries without a specific AI strategy at governmental level—such as Israel, Switzerland and the Russian Federation—the shading indicates the sectors in the country that have taken a lead in AI).

Germany

The federal government published a national strategy in November 2018, built around three main goals (German Federal Government, 2018, p. 8–9) :

- "We want to make Germany and Europe a leading centre for AI and thus help safeguard Germany's **competitiveness** in the future."
- "We want a responsible development and use of **AI which serves the good of society**."
- "We will integrate AI in society in ethical, legal, cultural and institutional terms in the context of a **broad societal dialogue** and active political measures."

The strategy aims to develop AI to solve specific problems with a great emphasis on machine learning, deduction systems, and human-machine interaction (ibid., p. 5). It presents 12 priority fields of actions, covering most of the domains presented in Table 4 above:

- Strengthening research in Germany and Europe
- Innovation competitions and European innovation clusters
- Transfer to business, strengthening the Mittelstand
- Fostering new businesses and leading them to success
- Shaping structural labour market change
- Strengthening vocational training and attracting skilled labour

- Using AI for public administration
- Making data available and facilitating their use
- Adapting the regulatory framework
- Setting standards
- National and international networking
- Engaging in societal dialogue about policy in this area

Specific attention is paid to the ethical considerations raised by “autonomous machinery and vehicles”, and an “ethics by design” perspective is advocated (ibid., p. 39). On the labour market implications of AI/ML, the strategy stresses the need for “a joint definition of the goals for ‘decent work by design’, which will be further specified and integrated into a ‘digital bill of rights’” (ibid., p. 39).

The Data Ethics Commission actively advises and supports the Federal Government. In October 2019, it produced an Opinion on data and algorithmic systems. The Opinion includes a recommendation to adopt a "Criticality pyramid and risk-adapted regulatory system for the use of algorithmic systems". It notes the need to differentiate data governance and governance of algorithms. The Commission stated its belief that "the state has a particular responsibility to develop and enforce ethical benchmarks for the digital sphere that reflect this value system, and that excessive dependence on others turns a nation into a rule taker rather than a rule maker, e.g. by private corporations that are exempt from democratic legitimacy and oversight." This statement is in line with recommendations that later chapters in this report will suggest. The Commission also "holds the view that regulation is necessary, and cannot be replaced by ethical principles – but also states that not everything that is relevant from an ethical perspective can and should be enshrined in legislation. Regulation must not unduly inhibit technological and social innovation and dynamic market growth." This view is also reflected in this report's recommendations. (Tranberg, 2019)

Germany’s strategy is supported by funding from, among others, the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung; BMBF). The BMBF has funded applied AI since 1988 its current annual investment in this area is EUR 40-50 million. This is separate from the budget for the German Research Centre for AI (DFKI), which received EUR 200 million (Stix, 2018). Other important features of the German AI/ML ecosystem include a number of strong and competitive academic institutions (including the Technische Universität München-TUM, the Max-Planck-Institutes for Intelligent Systems, and the Cyber Valley initiative funded by the state of Baden-Württemberg), industry players at global level (including Siemens, Mercedes, Volkswagen, BMW, Bosch, SAP, and Telekom) as well as the country’s start-up ecosystem.

United Kingdom

Since 2017 a number of key reports, policy documents, industry white papers (see Clark, 2017) and recommendations from academic independent research (see Hall & Jérôme, 2017) have established the framework for the development of AI/ML in the UK (Stix, 2018). An important outcome of this process has been the creation of the Alan Turing Institute—the national institute for artificial intelligence and data science—and the creation of a Centre for Ethics and Innovation which aims to make the benefits of AI/ML accessible and inclusive (Greg et al., 2019). In 2019, a Government Office for Artificial Intelligence was announced, as well as a new AI Council of experts to advise it.

The government's broad industrial strategy includes AI as one of four "grand challenges" (the others being future of mobility, clean growth, ageing society) and sets out five foundations for productivity: ideas, people, infrastructure, business environment, places (Greg et al., 2019). Particular emphasis is placed on machine learning technologies in relation to the enhancement of the existing data infrastructure. There is also a specific requirement that "industry will work closely with the government, through the AI Council, on broader questions related to AI such as data ethics and the role of AI in the public sector" (Greg et al., 2019).

The UK's strategy envisages close cooperation between government and industry, with an important role for the Engineering and Physical Sciences Research Council (EPSRC), the UK's main funding body for research in engineering and the physical sciences. Finally, it is worth noting the close links between the UK's AI ecosystem with other countries, such as the US and Japan, including investments from major players such as Google, Element AI, Amazon or Astroscale (Greg et al., 2019).

France

France began positioning itself on AI/ML in 2017 with the initiative "France Intelligence Artificielle: La stratégie IA en France." This was followed by a national strategy entitled "AI for Humanity" (https://www.aiforhumanity.fr/en/) and, in 2018, an influential report written by the mathematician Cédric Villani entitled "For a Meaningful Artificial Intelligence". The Villani report places great emphasis on machine learning, and identifies the following six key objectives (Villani, 2018):

- Building a data-focused economic policy
- Promoting agile and enabling research
- Assessing the effects of AI on the future of work and the labour market, and experiment adequate policy responses
- Artificial intelligence working for a more ecological economy
- Ethical considerations of AI
- Inclusive and diverse AI

On the basis of the Villani report, the following seven key themes were identified for the development of AI in France, each complemented by concrete proposals:

- Developing an aggressive data policy
- Targeting four strategic sectors: health, transport, environment and defence
- Boosting the potential of French research
- Planning for the impact of AI on labour
- Making AI more environmentally friendly
- Opening up AI "black boxes"
- Ensuring that AI supports inclusivity and diversity

It is also worth noting an early initiative from the National Commission for Information Technology and Liberties (CNIL), which organised a public debate and produced a report about moving from ethical thinking to algorithmic regulation (Demiaux & Abdallah, 2017). The report raised questions on issues such as the potential prohibition of AI in certain sectors, such as medicine, justice or the military (ibid., p. 47). A set of principles have been formulated including the principle of fairness, the principle of continued attention and vigilance as well as the engineering principles of intelligibility, accountability, and human intervention for AI systems. These principles have eventually led to the following recommendations:

1. "Fostering education of all players involved in the 'algorithmic chain' (designers, professionals, citizens) in the subject of ethics"
2. "Make algorithmic systems understandable by strengthening existing rights and organising mediation with users"
3. "Improve the design of algorithmic systems in the interests of human freedom"
4. "Set up a national platform for auditing algorithms"
5. "Increasing incentives for research on ethical AI and launching a participatory national worthy cause on a general interest research project"
6. "Strengthen ethics within businesses"

Other initiatives in France have included the launch of new incubators and funding for AI. In May 2018, President Emmanuel Macron, while presenting the French vision and strategy for AI, announced EUR 1.5 billion of investment into AI research (Future of Life Institute, 2019d).

Italy

Italy's AI environment includes recognized research expertise, industrial initiatives (with a strong focus on the delivery of machine learning-related services) and a start-up ecosystem (OECD, 2017a) (Stix, 2018). In 2017 an AI Task Force was created through the Agency for Digital Italy (AGID). Mainly focused on the use of AI in public administration, the task force produced a white paper in 2018 entitled "AI at the service of citizens" which listed the following priorities:

- ethics

- technology
- skills
- the role of data
- the legal context
- accompanying the transformation
- preventing inequalities
- measuring impact
- the human being

With respect to ethics, the task force notes in particular issues including: data quality, errors and biases in machine learning systems, accountability and liability, transparency and openness, and privacy. Three of the task force's final recommendations are worth mentioning:

- the disclosure to the public of "the intermediate results of the elaboration of AI algorithms (for example, the parameters of neural networks) operated on data from public administrations, subject to conditions that may harm the privacy and security of citizens" (AItaskforce, 2018, p. 69);
- the development of resources for computational linguistic systems focused on the Italian language;
- the development of "adaptive customization and recommendation systems that facilitate interaction with the services offered by public administrations based on the specific requirements, needs and characteristics of citizens" (Altaskforce, 2018).

The white paper also recommends the establishment of "security-by-design" guidelines and processes to facilitate cooperation and data-sharing among European countries to counter cyber-attacks (AItaskforce, 2018). Another initiative of the AI Task Force is the Observatory on Artificial Intelligence, which tracks AI-related public conversations on social networks.

Denmark

Denmark has one of the more structured strategies on AI, supported by an extensive plan of investments. The country published its national strategy for AI in March 2019 (Danish Government, 2019). Its overarching goal is for Denmark "to be a front-runner in responsible development and use of artificial intelligence" (ibid., p.7), and to this end it sets out four main objectives (ibid., p. 8-10):

- a common ethical and human centered basis for AI
- AI research and development (R&D) artificial intelligence
- growth through the development and use of AI by businesses
- the use of AI to deliver world class public-sector services

The strategy identifies the following core values or principles: self-determination, dignity, responsibility, explainability, equality and justice, and development (Danish Government, 2019). Furthermore, the actions recommended by the strategy are structured around four focus areas where intervention is prioritised:

- a responsible foundation for artificial intelligence
- more and better data
- skills and knowledge
- increased investment.

These four focus areas incorporate twenty initiatives, and the Danish strategy highlights a number of priority domains where specific investments have been allocated. These investments are summarised below in Table 5. Finally, the Danish government has also planned a set of signature AI-related projects in areas including health, social services and employment (ibid., p. 21).

Table 5: Investments for areas/initiatives from Danish national strategy for AI

Areas/Initiatives	Investments
Public research budget	Budget 2019: DKK 23 billion (EUR 3.1 billion).
Cyber and information security	Ministry of Defence for the years to come: DKK 1.5 billion (EUR 200 million).
Health Data Programme (data quality and cross-sectoral cooperation on health data)	The government has earmarked DKK 250 million.
Skills (in all technical fields and new technologies, including artificial intelligence)	A pool of DKK 190 million (EUR 25 million) by the government.
Artificial intelligence in the public sector (with **special attention to municipalities and regions).**	The fund will have a total investment budget of DKK 410 million (EUR 55 million) up to 2022.
Technological services for Danish businesses	More than DKK 600 million (EUR 80.5 million) for the period 2019-2020.
Fuel-efficient routes project (algorithms that can manage systems associated with great uncertainty)	Independent Research Fund Denmark has granted DKK 5.8 million (EUR 800,000) .

Source: (Danish Government, 2019)

Sweden

Sweden adopted its AI strategy—the "National Approach to Artificial Intelligence"—in 2018 (Ministry of Enterprise and Innovation, 2018). The strategy identifies four key conditions for the

use of AI in Sweden, and stresses the importance of including all sectors of societies in reaching pursuing these:

- education and training
- research
- innovation and use,
- framework and infrastructure.

In view of the discussion in section 4.1, it is worth noting that the Swedish strategy highlights the significance of targeting machine-learning research in order to compete in AI leadership with countries such as the United States and China (ibid., p.7).

The strategy calls for "pilot projects, testbeds and other specialised testing environments" in order to accelerate the introduction of "new AI technology in an ethical, safe, secure and sustainable manner". It also highlights the important role of EU regulatory frameworks—and the GDPR in particular—in determining "how well Sweden is able to manage both the benefits and risks of AI" (ibid., p. 10).

In addition to Sweden's existing capabilities in computer science and mathematics a number of new investments have been ear-marked for the development of AI. These include the donation of the SEK 1 billion to AI research in November 2017 by the Knut and Alice Wallenberg Foundation (KAW), as well as the SEK 40 million ($4.2 million) invested by the Swedish government in AI training (ibid., p. 7; Moltzau, 2019).

Finland

In 2017 the Ministry of Economic Affairs and Employment published an initial report on AI (Ministry of Economic Affairs and Employment, 2017), which highlighted the potential to use AI to:

- increase business competitiveness,
- deliver high-quality public services,
- improve the well-being of citizens.

The report highlighted the need to develop skills in machine learning technology in academic curricula. A period of public debate was followed in 2019 by the publication of a final AI strategy (Ministry of Economic Affairs and Employment, 2019), which identified the following eleven priorities, setting goals and timetables for each:

- "Enhance business competitiveness through the use of AI"
- "Effectively utilise data in all sectors"
- "Ensure AI can be adopted more quickly and easily"
- "Ensure top-level expertise and attract top experts"
- "Make bold decisions and investments"
- "Build the world's best public services"

- "Establish new models for collaboration"
- "Make Finland a forerunner in the age of artificial intelligence"
- "Prepare for artificial intelligence to change the nature of work"
- "Steer AI development into a trust-based, human-centred direction"
- "Prepare for security challenges"

The strategy incorporates a range of concrete actions, some of which had already been put in place under pre-existing initiatives by companies including Cargotech, KONE as well as startups such as ultimate.ai. Other initiatives in Finland include the Finnish Centre for AI, a partnership between Aalto and Helsinki Universities which focuses on research, talent development and collaboration with industry (Moltzau, 2019).

Estonia

Estonia is a leader in digitalization, especially for the provision of public services and digital government. In June 2019 the government published its national strategy for accelerating the development of AI, named KRATT (Sikkut, 2019) (e-estonia, 2019)[17]. The strategy envisages the public sector having a key role in driving the development of AI, with a particular focus on ensuring data access and quality. However, notwithstanding the strong public-sector focus, pilot programmes are expected to fully consider business needs as well as those of researchers and innovators. The strategy does not call for significant regulatory change in order to adapt to AI, stating that there is "no need for changes in the foundation of the legal system and there is no need for a unified AI law" (e-estonia, 2019). According to the official e-estonia portal (2019), the following issues will need to be taken into account to accelerate the implementation of the strategy:

- "Open source base components and other 'tools'"
- "Data science and AI collaboration network within the public sector"
- "Guidelines on how to manage AI projects (including sustainable development)"
- "Knowledge transfer and experience exchange about projects and possibilities across different networks and in different formats"
- "For the sustainable development of Estonian language and culture it is important to use natural language processing"
- "New positions of Chief Data Officers at least on Ministerial levels"
- "Flexibility in funding e-state developments, so that there would be enough resources for developing and testing "Kratt" applications"
- "Create 'sandboxes' for testing and developing public sector solutions"
- "Technical requirements as part of funding conditions for AI projects to ensure longevity"

[17] The strategy document is only available in Estonian; our summary draws on the discussion of the strategy on the official e-estonia portal (2019).

- "Data governance workshops and data auditing"
- "'Bureau-Kratt' – concept of a personal virtual assistant through collaboration across networks"

Switzerland

Switzerland presents something of a paradox in relation to its positioning on AI. On the one hand, it does not yet have a specific national strategy for AI, and takes what has been described as a "passive stance" towards AI (pwc, 2019). On the other hand, the country is already an important global player in a number of respects. It conducts world-class research (with ETH Zurich and EPFL among the top global players), and the corporate sector is investing in AI and developing new research centres. The level of new ventures in the AI area is also increasing, positioning Switzerland as a "hub for artificial intelligence"(Donatelli, 2018).

Israel

Israel is in a similar position to Switzerland. It has a vibrant AI ecosystem of research, innovation and industry, but the government has yet to finalise its work on a national strategy "to make the country a leader in artificial intelligence" (Kelly, 2019). In Israel's case, the strength of the ecosystem may explain why the government is lagging behind others in producing strategic plans and initiatives. In the words of Ben-Israel, a key figure in the Israeli research community and a leader of work on the government's AI vision: "There's not much that needs to be done, in fact; the government has to send a signal, that's all […] There is in fact a lot of money already invested in AI research here by global companies. Nobody waits for the government to do something. It happens automatically." (Kelly, 2019). Israel's ecosystem of AI startups totals more than 1,000 companies (NoCamels Team, 2019), with 51% of them using machine-learning technologies in particular (Singer, 2018). Major global players are present, including IBM, Intel, Microsoft and Google. Israel is also active in international research networks and is represented in various research projects funded by Horizon 2020.

Saudi Arabia and the United Arab Emirates (UAE)

The United Arab Emirates was the first country to appoint a government minister dedicated to AI, in October 2017 (DUBAI FDI, 2018). It was also among the first countries to put in place an AI strategy. The UAE strategy focuses on using AI to "boost performance" in the following sectors in particular:

- transport: to reduce accidents and cut operational costs
- health: to minimise chronic and dangerous diseases
- space: to help conduct accurate experiments, reduce rate of costly mistakes
- renewable energy: to manage facilities

- water: to conduct analysis and studies to provide water sources
- technology: to increase productivity and help with general spending
- education: to cut costs and enhance desire for education
- environment:to increase forestation rate
- traffic: to reduce accidents and traffic jams and draw more effective traffic policies

In Saudi Arabia the Vision 2030 and National Transformation Programme 2020 identify AI-based technologies as key instruments for improving healthcare, for helping citizens face future labour market patterns, and for developing new public-private business models to reduce the country's dependence on oil revenues (Jain et al., 2018). In 2017, Saudi Arabia was the first country to grant citizenship to a humanoid robot, produced by Hanson Robotics.(Future of Life Institute, 2019c).

Canada

Canada enjoys a favourable environment for AI, with "skilled workforces, innovative private sectors, good data availability, and effective governance" (Miller et al., 2019). AI is a national policy priority area, and the country launched an AI strategy in 2017 with a specific focus on human capital and attracting highly-skilled AI talent (Miller et al., 2019). The government adopted a crowdsourcing approach (Afuah et al., 2018; Brabham, 2013; Howe, 2006) when producing the white paper "Responsible Artificial Intelligence in the Government of Canada" [18]. This was followed in 2019 by the publication of a "Directive on Automated Decision-Making" (OPSI, 2019), which identifies conditions of transparency, accountability, legality and fairness that must be met when the public sector wants to use public-facing decision-making algorithms (the directive does not cover internal government services or national security issues). The most governance instrument in the directive is an Algorithmic Impact Assessment, a digital questionnaire that agency leaders have to complete "before producing or significantly changing an automated decision system" (OPSI, 2019). Canada is also using a range of other tools and incentives to shape the governance of AI, such as the creation of an AI Source List with 73 pre-approved suppliers "to provide Canada with responsible and effective AI services, solutions and products" (OPSI, 2019).

India

The country published an AI strategy in June 2018 (Kumar et al., 2018), which argued for an inclusive approach named #AIFORALL and focused on "empowering human capability and ensuring social and inclusive growth" (OECD, 2019a, p. 130). Among the specific AI applications highlighted are healthcare, agriculture, education, smart cities and transportation. One of the

[18] Available at https://docs.google.com/document/d/1Sn-qBZUXEUG4dVk909eSg5qvfbpNlRhzIefWPtBwbxY/edit [Accessed on September 18, 2019]

challenges identified by the strategy is the need to create data ecosystems. (The development of repositories of healthcare data for machine learning applications has been cited as a priority in India's healthcare system (Saran, Natarajan, & Srikumar, 2018, p. 35).) In the area of data protection, the EU's governance framework has been a prominent reference point in India's ongoing development of its own regulatory framework. An initial draft data protection bill was published in 2018, drawing on important aspects of the GDPR. However, subsequent revisions have significantly altered the bill, in ways that chair of the committee that drafted the original bill has described as "dangerous" and "Orwellian".

Japan

In 2017 Japan published the document "Draft AI R&D GUIDELINES for International Discussions", with a view to sharing a set of proposed guidelines that could be used internationally as the basis for "non-regulatory and non-binding soft law" (The Conference toward AI Network Society, 2017, p. 2). The proposed guidelines comprise five basic principles:

- the achievement of a human-centered society
- non-binding soft law and best practices
- ensuring an appropriate balance between the benefits and risks of AI networks
- technological neutrality (not imposing excessive burden on developers)
- constant update and renew of the guidelines

Furthermore, the document established nine principles to guide R&D related to AI:

- transparency
- controllability
- safety
- security
- privacy
- ethics
- user assistance
- accountability

Also in 2017, the Strategic Council for AI Technology published a roadmap that included the following priority areas (Strategic Council for AI Technology, 2017, p. 4):

- productivity
- health, medical care, and welfare
- mobility
- solutions for social issues
- contribution to economic ripple effects
- information security".

Japan has also enforced recommendations and guidelines for developers, to mitigate risks (particularly relating to individual rights) “arising from the interconnectivity and interoperability of AI systems”. These guidelines suggest the use of controlled environments and “sandboxes” and emphasise the principle of controllability (Saran et al., 2018; The Conference toward AI Network Society, 2017).

South Korea

In addition to hosting one of the key corporations for digital transformation, Samsung, the government of South Korea has created an institutional framework that is designed to increase the role of the country in the global industrial ecosystem. In March 2016 it published the “Intelligent Information Industry Development Strategy”, announcing public investment in AI and related information technologies totalling KRW 1 trillion (USD 940 million) by 2020 (OECD, 2019a). This was followed in December 2016 with the publication of the “Mid- to Long-Term Master Plan in Preparation for the Intelligence Information Society”. This strategic plan focused on so-called “intelligent IT”, defined as “technology that is capable of performing the highly complex functions of human intelligence by combining the ‘intelligence’ of artificial intelligence (AI) with the ‘information’ provided by data-processing and network technologies” (MSIP, 2016, p. 6). The plan defined a vision of a “human-centered intelligent information society” and identified the following core factors for success (MSIP, 2016):

- “Enhancing Korea’s Intelligent IT capabilities and strengthening its data infrastructure”
- “Achieving greater convergence between Intelligent IT and existing industries”
- “Reforming the labor market and increasing education for creative personnel”

The plan also outlines a set of policy tasks. One example is its call for the establishment of “a national data management system for the development of large scale data infrastructure that facilitates machine learning”. In order to create this national data management system, public bodies (starting with 20 in 2018, and covering all bodies by 2025) to identify the high-utility data they rely on (for example, data relating to medicine, patents, and languages) (MSIP, 2016).

South Korea’s strategic plancalls for the creation of an AI-based administrative service system capable of responding automatically to service requests from citizens (ibid., p. 41). It also anticipates significant legal implications relating to AI, suggesting the need “to grant rights and responsibilities to ‘electronic persons’ in preparation for the dissemination of AI and self-learning machines” (ibid., p. 56). The plan stresses the need for systems to evaluate whether “secure and appropriate data have been input into the machine learning process of Intelligent IT software, and whether such software is capable of quickly identifying and resolving errors” (ibid., p. 59). The plan also highlights security issues, noting the need for AI security professionals, as well as the enhancement of global partnerships and the establishment of cyber-security and machine-learning training facilities and incentives (ibid., p. 59).

Other initiatives include plans to boost inclusion, such as the use of machine learning-based mobility aids, and systems for people with disabilities (Ibid., p. 55). The plan also suggests the establishment of a data environment for effective and quality healthcare services—for example, the development of a comprehensive AI-based health database that would integrate data (including electronic medical records and genetic information) from multiple sources, including hospitals, the National Health Insurance Service (NHIS) and the Health Insurance Review and Assessment Service (HIRA) (ibid., p. 46).

Finally, it is worth noting that Korea and the EU have jointly begun a 90 billion research program across the entire research and development cycle with joint commercialisation goals (Saran et al., 2018).

Australia

The Australian government allocated AUD 28 million (USD 21 million) in its 2018/19 budget for AI and machine learning development, with a focus on education, research, standards labour-force impacts and an AI ethics framework. (Future of Life Institute, 2019a; OECD, 2019a). Australia's ethics framework is based on case studies. It focuses on three key aspects of AI/ML, and highlights the principles that are at stake in each case (Dawson et al., 2019, p. 5):

- data governance and AI (privacy and fairness)
- automated decision-making (fairness, transparency, explainability, contestability, accountability)
- prediction of human behaviour (do no harm, regulatory and legal compliance, privacy, fairness, transparency and explainability)

The Australian government is still working on a national strategy for AI. It is being urged on in this endeavour by the Australian Council of Learned Academies (ACOLA), which in 2019 published a report on the "effective and ethical development of artificial intellingence" (Elliott, 2019; Walsh et al., 2019). The ACOLA report notes the potential for AI to improve Australia's economic, societal, and environmental wellbeing, but points out the need for a balance between the benefits of innovation (for example on the infrastructure and economy) and the potential associated risks (such as labour-market disruption or the development of "autonomous weapons") (Elliott, 2019).

Among other initiatives, it is worth noting the proposal of Australia's chief scientist for an AI certification that companies would receive for meeting "ethical standards and independent auditing requirements for AI development" (Clark, 2018; Future of Life Institute, 2019a).

Russian Federation

The geopolitical importance attached to the development of AI in the Russian Federation was highlighted in 2017 when President Putin stated that "whoever becomes the leader in this sphere

will become the ruler of the world". A number of policy initiatives have subsequently been launched, such as a ten-point plan for AI development in Russia published in 2018 by the Ministry of Defense, the Ministry of Education and Science, and the Academy of Sciences. The document covers, among other things, funding, educational plans, as well as the organization of AI war games (Bendett, 2018; Future of Life Institute, 2019b). Russia supports the development of AI for military use and is strongly opposed to the prohibition of “lethal autonomous weapon systems or LAWS” (Hutchison, 2017). The 2018 ten-point plan was followed in 2019 by the development of a national strategy for AI based on the same principles and on President Putin’s vision for the technology. In May 2019 (Cress, 2019) the president used another speech to highlight some of the key issues in the strategy, including:

- education and training in STEM subjects
- research and development
- legislative support for AI standards
- workforce creation
- investment in a digital national technology program (Bendett, 2019; "Russia’s National AI Strategy Takes Shape", 2019).

It is also worth noting plans for public-private partnership in the creation of an “AI infrastructure” through a deepening of relationships between the country’s state and private high-tech sectors (Bendett, 2019; Cyber Security Intelligence, 2019).

4.3. Intergovernmental organizations

Intergovernmental and other international organisations offer diverse and multiple AI initiatives. We will review a range of these, with a particular focus on the Organisation for Economic Co-operation and Development (OCED).

OECD

The OECD is an influential actor in the emerging global governance system for artificial intelligence and machine learning. To understand the evolution of the OECD's work on AI and its impact on societies and economies, it is useful to consider first the proceedings of a 2016 “Technology Foresight Forum” on the economic and social implications of artificial intelligence” (OECD, 2017b), at which participants discussed the challenges faced by different actors at both national and global levels. A number of important issues were highlighted:

- avoiding hype and “sensationalism” and to focus on evidence-based communication about AI
- focusing on domain-specific applications and challenges so as to avoid trying to grapple with “the inconceivably vast number of things an intelligent system might think or do”
- preparing societies for the changes—both opportunities and risks—that AI will bring

- setting limits for automated decision-making
- minimizing risks related to “discrimination, privacy erosion, loss of public anonymity”
- implementing "ethics by design"
- respecting the rules of societies and recognising the different perspectives on fundamental human rights and values that exist among the global players
- harnessing the predictive power of AI/ML but maintaining vigilance for biases that can enter the process at various stages and that pose a danger to civil liberties in applications such as predictive policing or facial recognition
- preventing excessive increases in income inequality
- the implications of a "winner-takes-all" dynamic, potentially concentrating of computational and market power in a small number of companies and governments
- possible societal responses to various AI risks, from universal basic income to the free or open source distribution of machine learning technology
- increasing the explanability, transparency and understandability of algorithms
- promoting education
- ensuring that regulatory frameworks uphold the principles of responsibility, liability, security and safety (OECD, 2017b).

“OECD AI Principles” were accepted by all OECD member countries in May 2019 (OECD, 2019c), and informed the “human-centred AI principles” that the G20 group of countries approved in June 2019 (OECD, 2019b). The five OECD principles are:

- AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being
- AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards—for example, enabling human intervention where necessary—to ensure a fair and just society
- There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them
- AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed
- Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles

Another OECD initative on AI is the creation of the OECD Artificial Intelligence Policy Observatory, launched in February 2020. The Observatory provides data and multi-disciplinary analysis on

artificial intelligence, with a view to supporting governments' efforts to develop public policies for responsible, trustworthy and benefical AI. [19]

In 2019 the OECD's Observatory of Public Sector Innovation (OPSI) initiative published a draft primer for public servants aiming to support the use of **AI/ML in driving public-sector innovation and transformation** (OPSI, 2019). The document discusses the main issues related to AI/ML and in the public sector and presents a series of case studies from around the world. The OPSI primer also sets out a three-stage process for governments seeking to deploy AI in the public sector:

- **Establishing a baseline:** "an assessment of the organisation's current strategic situation and challenges that AI might help address"
- **Setting objectives:** "what the organisation wants to achieve using AI and the principles that will underpin the actions it takes to achieve them"
- **Deciding on approaches:** "the concrete actions that will be undertaken to achieve these objectives", such as ensuring access to necessary skills or data, and putting in place legal, ethical and technical frameworks

International Telecommunication Union (ITU)

In 2017 the International Telecommunication Union (ITU) co-organised (with other UN agencies, the XPRIZE Foundation and the Association for Computing Machinery) a global summit on "AI for Good". The summit emphasised the need for developing and deploying machine learning for the greater good rather than for narrower self-interest. It stressed the need for institutional solutions to facilitate knowledge-sharing and collaboration, and pointed to collaboration at the European Organization for Nuclear Research (CERN) as a model (ITU, 2018). ITU has also launched an AI repository, with a view to identifying AI-related projects, research initiatives, think-tanks and organizations that can accelerate progress towards the Sustainable Development Goals (SDGs) (ITU, 2018).

In 2018, ITU also partnered with the World Health Organization (WHO) to create a joint Focus Group on artificial intelligence for health (FG-AI4H), tasked with establishing "a standardized assessment framework for the evaluation of AI-based methods for health, diagnosis, triage or treatment decisions" (ITU/WHO, 2018). The group has produced, among other outputs, a set of policies on data acceptance and data handling, a thematic classification scheme, and a white paper that emphasises the need for increased interpretability, explainability, and proven robustness of deep learning models in order to improve their societal acceptance "when facing critical or even vital decisions" (Salathé et al., 2018, p. 3). The white paper also highlights

[19] See https://oecd.ai/

implications of data protection laws for access to the training data needed to improve the predictive performance of healthcare machine-learning models (Salathé et al., 2018).

International Panel on Artificial Intelligence (IPAI) and Global Partnership for AI (GPAI)

In 2019, the French and Canadian governments published a Declaration of the International Panel on Artificial Intelligence.[20] The intention is that the IPAI will serve as a disinterested source of global expertise on the societal issues raised by AI—it is modelled on the highly influential Intergovernmental Panel on Climate Change (IPCC). (Nature, 2019). In November 2019 France and Canada moved forward again with plans for a 'global AI expert council', that would create a 'Global Partnership on AI' (GPAI), with the OECD joining the two countries for policy discussions around the creation of a standing forum – involving government, industry and academia – to monitor and debate the policy implications of AI globally.

4.4. Other institutions

In this section, we review a selection of AI governance-related initiatives that have emerged from the scientific, academic and non-profit sectors. We have not included proposals on AI governance developed by individual private-sector actors. This is a burgeoning area of activity, which goes beyond the scope of this report.

Institute of Electrical and Electronics Engineers (IEEE)

The Institute of Electrical and Electronics Engineers (IEEE), a standard-setting association, has published a series of documents as part of an initiative on the “Ethics of Autonomous and Intelligent Systems” (2018). One of the IEEE’s main objectives with this initiative is to prompt a public discussion about the ethical, social and cultural considerations involved in deploying AI/ML in a way that protects “well-being”. The initiative is also designed to develop a set of standards in this area (the IEEE P7000TM series), as well as associated certification programs, and finally to enable the creation of a policy framework.

It is worth noting that the issues and arguments raised by the IEEE have been found in quite a few other reports considered in the course of our analysis, produced by institutions such as the The National Academy of Sciences and The Royal Society (National Academy of Sciences, 2018) or the Future of Humanity Institute (Brundage et al., 2018).

The IEEE focuses on the broad category of “autonomous and intelligent systems (or A/IS)”. This includes multiple fields and is framed in this way to ensure the “broadest application of ethical considerations in the design of these technologies as possible.” (ibid., p. 12). With respect to

[20] Canada, France Governments Announce Declaration of the International Panel on AI (Government of Canada, 2019)

machine-learning more narrowly, a key challenge highlighted by the IEEE relates to “evaluation by third parties”, which it notes is complicated by factors including:

- inaccessibility of data sets to evaluators
- proprietary algorithms
- the opacity of the specifications of the final systems
- mathematical complexity and the frequent use of “black-box” components, and (ibid., p.52) (ibid., p. 70).

The IEEE stresses that such issues should be taken into account in order to enable society to maximise the potential benefits of widespread deployment of machine learning. In relation to automated decision-making, it argues that as long as appropriate protections are put in place, “we must assume that virtually every decision that we make as humans can be mediated or replaced by an algorithm” (ibid., p. 158). To take a particular example, it claims that the use of machine learning to scan resumés/CVs during the recruitment process can lead to increased fairness and reduced biased, “provided that the systems are designed well” (ibid., p. 158).

The IEEE is not the only standards-setting body that can be expected to address the challenges posed by artificial intelligence. The US and Chinese governments have both put their weight behind the creation of international standards, and the ISO/IEC Joint Technical Committee for Information Technology has begun looking at artificial intelligence. (Cihon, 2019)

World Economic Forum (WEF)

Various initiatives and publications by the World Economic Forum have emphasized the twin benefits from AI and machine learning of boosting economic growth (particularly in sectors such as automotive, logistics and retail) and of creating value for society (particularly in healthcare, through the introduction of new treatments) (World Economic Forum, 2019).

The WEF has sought to outline the potential challenges and risks associated with AI/ML, particularly with respect to issues of complexity, opaqueness, ubiquity and exclusiveness (Bernstein et al., 2018, p. 7). This is part of an effort to provide a framework for organizations to better understand the implications of AI/ML. The table below outlines a number of key questions that need to be answered, according to the WEF. In a 2018 white paper entitled “How to Prevent Discriminatory Outcomes in Machine Learning” (Bernstein et al., 2018), the WEF draws on the Universal Declaration of Human Rights to present four central principles against bias in machine learning. These are active inclusion, fairness, right to understand, and access to redress.

Table 6: Issues and questions raised by AI and machine learning, adapted from World Economic Forum (2019).

Issues	Questions

Unemployment	*"As machines take over more mundane tasks, how do we prepare displaced humans to fill the roles created by new technologies?"*
Inequality	*"How should the wealth created by machines be distributed?"*
Humanity	*"How do interactions with machines affect our behaviour?"*
Artificial stupidity	*"How can we guard against mistakes?"*
Racist robots	*"How do we eliminate AI bias?"*
Security	*"How do we keep AI safe from adversaries?"*
Evil genies	*"How do we protect against unintended consequences?"*
Singularity	*"How do we maintain control over complex, intelligent systems?"*
Robot rights	*"How do we treat AI humanely"*

AI Now Institute

The AI Now Institute at New York University (NYU) is dedicated to understanding the social implications of artificial intelligence, with a focus on labour, automation, bias, inclusion, safety and critical infrastructure (AI Now, 2019). Its work includes a framework for algorithmic impact assessment for public agencies (Reisman et al., 2018), and a policy toolkit that illustrates core AI concepts and provides information on AI products and vendors (AI Now, 2018).

In its annual report, the AI Now Institute provides a set of cases and supporting data with a focus on the following problems relating to the deployment of AI:

- an AI accountability gap
- an increase in surveillance
- the testing of AI systems in public spaces
- a growing tendency by governments to use automated decision systems "without adequate protections for civil rights" (Whittaker et al., 2018, p. 8)

In addition, the AI Now Institute questions the focus on designing mathematical models to deal with values such as fairness, without taking sufficient account of "social and political contexts and histories" (ibid., p. 8). It also notes the limitations of focusing on ethical principles if these are not "directly tied to structures of accountability" and "backed by enforceable mechanisms of responsibility that are accountable to the public interest" (ibid., p.9).

Nuffield Foundation

An effort to organize the growing body of work on the values implications of AI/ML has been carried out by researchers from the Leverhulme Centre for the Future of Intelligence for the Nuffield Foundation (Whittlestone et al., 2019, p. 2). Framed as a "roadmap for work on the ethical and societal implications of algorithms, data, and AI" the Nuffield report reviews commonly used ethical concepts and highlights four central tensions (summarised in Table 7 below) that may arise with the use of AI/ML. If left unchecked, these tensions may eventually result in self-

reinforcing cycles of discriminatory outputs and social injustices, the report states. To counter this risk, the report recommends focusing research and policy work on building a better evidence base relating to technological capabilities and limitations, and to societal attitudes and needs (ibid., p. 35).

Table 7: Four central tensions between technology goods and values classified per type of value. Adapted and elaborated from Whittlestone et al. (2019)

Tensions	Technology Goods	Values	Type of value
"Using algorithms to make decisions and predictions more accurate versus ensuring fair and equal treatment."	Accuracy	Fairness	Societal values
"Reaping the benefits of increased personalisation in the digital sphere versus enhancing solidarity and citizenship."	Personalisation	Solidarity	
"Using data to improve the quality and efficiency of services versus respecting the privacy and informational autonomy of individuals."	Quality and efficiency	Informational autonomy	Individual values
"Using automation to make people's lives more convenient versus promoting self-actualisation and dignity."	Convenience	Self-actualisation	

Harvard University (Berkman Klein Center)

Another initiative to map the work being done on ethical AI is under way at the Berkman Klein Center for Internet and Society at Harvard University. The project, "Principled Artificial Intelligence: Mapping Consensus and Divergence in Ethical and Rights-Based Approaches" (Fjeld et al., 2019), aims to produce a visualisation of based on a dataset comprising various sets of AI principles that have been produced by governments, private companies, intergovernmental organizations, multi-stakeholder projects and civil society. The project seeks to highlight points of commonality and divergence in the dataset. It identifies the following eight shared themes across them:

- Accountability
- Fairness and non-discrimination,
- Human control of technology

- Privacy
- Professional responsibility
- Promotion of human values,
- Safety and security, and
- Transparency and explainability

A second relevant project at the Berkman Klein Center ("Artificial Intelligence and Human Rights: Opportunities & Risks" (Raso et al., 2018)) seeks to evaluate the human rights impacts of six AI domains of application. This involves a detailed analysis of how specific AI technologies (such as machine learning) are implemented, in order to gauge the impact on human rights (with a particular focus on international law). The project covers the following six AI domains, the first two of which overlap with the domains covered in chapter 3 of this report:

- Criminal Justice (*risk assessments*)
- Healthcare (*diagnostics*)
- Finance (*credit scores*)
- Content Moderation (*standards enforcement*)
- Human Resources (*recruitment and hiring*)
- Education (*essay scoring*)

4.5. Conclusion

The foregoing sections have covered a diverse range of strategies, initiatives and projects. Although this chapter is not designed to be comprehensive, it demonstrates the volume and variety of activity in this area. While noting the danger of over-simplifying the heterogeneity of the sample we have reviewed, there are a number of recurring themes that can be noted in conclusion. These can be seen as important "landmarks" in the governance landscape, of which any successful governance initiative in the EU or elsewhere will have to take account:

- The potential benefits of AI/ML are widely acknowledged, especially in terms of the impact on economic growth (Purdy & Daugherty, 2016) increased productivity and operational efficiency.
- Another common strand is the need for mechanisms to provide for fairness and accountability—this is seen as an important factor in societal acceptance of AI/ML systems and applications (Smallman, 2019).
- More generally, many of the initiatives considered in this chapter note the difficulty of resolving the tensions between the various ethical values that can come into play when AI/ML is used (as argued by Whittlestone et al., 2019).
- When considering government strategies, it is broadly the case that in the EU normative considerations are given greater weight than in the US or China, where technological and economic development are less constrained.

- Distinctive aspects of the EU's approach to AI/ML include the attention paid to quality of life and inclusivity, as well as the potential to drive innovation in the public sector.
- The EU is likely to be comfortable with the prospect of regulating AI/ML, having already adopted a position of global influence as a data-protection regulator. The white paper on AI published in February 2020 (European Commissison, 2020) points to movement in this direction. But it has yet to be demonstrated how this can be implemented
- Developing a regulatory framework at the EU level would help to align the initiatives and activities that are under way in the member states, thereby harmonising governance patterns across the bloc. It seems that member states expect the EU to adopt that kind of governance leadership role.
- It is worth recalling that Europe is not alone in considering regulatory interventions. For example, the "algorithmic impact assessment" proposed in Canada has the potential to influence governance initiatives more widely.
- At a technical level data quality was a recurring theme, with particular importance for reducing errors and biases.
- A related point concerns the risks associated with "off the shelf" AI/ML solutions which may create unforeseen problems if deployed in different contexts from where they were developed—for example, because the training data used is unrepresentative of conditions where the technology is being deployed.
- At a much broader level, another recurrent governance challenge relates to the "winner-takes-all" dynamics that characterize many digital technologies and that risk concentrating AI/ML power and influence in a small number of countries and companies.

5. Key themes: innovation, ethics and niche leadership

In the preceding chapters, we have set out some of the most important contours of the landscape that the European Union confronts with respect to "governance of and by" machine-learning technology and its applications. We have outlined the basic workings of machine learning (chapter 1). We have highlighted prominent "acceptability challenges" that this technology raises, some of which are particularly acute in a European context (chapter 2). We have stressed the inadequacy of a one-size-fits-all approach to machine-learning governance, illustrating with three case studies how distinct governance challenges arise in different domains where machine learning is deployed (chapter 3). And we have presented an overview of European and global initiatives relating to governance and machine learning (chapter 4).

Together, these first four chapters represent a snapshot of the current state of machine-learning technology and governance. In the chapters that follow, we adopt a more forward-looking stance, highlighting some of the opportunities, risks and trade-offs that Europe now faces and should address to recover, maintain or increase its level of influence over governance in this area. We will do this in four stages. In this chapter, we argue that the relative influence of different global players (notably Europe, the US and China) is being shaped by the balances they strike between viewing ML as a source of adverse societal disruption or of welcome innovation and growth. The underlying values that prevail in different countries and regions are an important constraint on the caution/innovation balances that each can strike, and later in this chapter 5 we will consider how this affects the opportunities and risks that different stakeholders face. In particular, we will suggest that Europe appears well placed to pursue an "ethics first" strategy, which may limit the overall reach of its machine-learning influence, but which may give it deep influence in a number of specific domains or niches. In chapter 6, we will look more closely at this idea of a European "global niche leadership", looking at how the benefits of such a strategy could be maximized and how the associated costs or constraints might be minimized. We conclude in chapter 7 with a list of recommendations for Europe.

5.1. Balancing innovation and ethical caution

As the discussion in chapter 4 has made clear, machine learning (and artificial intelligence more broadly) is becoming a prominent focus for policy-makers across the world. Broadly speaking, our survey of the governance landscape suggests that two main responses can be identified: on the one hand, viewing machine learning as a source of innovation, growth and competitiveness to be actively promoted, and on the other hand, viewing it as a source of potential societal disruption to be managed with caution. We broadly group considerations of machine-learning efficiency and performance under the "innovation, growth and competitiveness" category, while under the caution/precaution category we include normative factors discussed in chapter 2, such as privacy, transparency and human agency.

Innovation and caution are not mutually exclusive responses. On the contrary, both responses are evident in all/most of the countries we have surveyed, with policy-makers seeking both to harness the economic benefits that machine learning offers, while also examining what safeguards might be needed to prevent or minimize any adverse effects.

The global governance landscape for machine learning is not shaped by a winner-takes-all contest between growth and caution, but by the balances that different countries and regions are striking between the two responses. For example, we have seen in chapter 4 that Europe is actively seeking to leverage machine learning as a source of competitiveness, but also that relative to the US and China it is nevertheless more focused on addressing the technology's potential adverse societal implications. The different weightings that different societies afford to growth and precaution reflect, among other things, the important role played in governance by the underlying value systems that prevail in each political system. Ethical values influence all policy areas, but they are particularly prominent with respect to machine learning because of the fundamental normative issues that this technology raises: as we saw in chapter 2, ML immediately touches on ethically charged questions of societal acceptability, such as safety, privacy, human agency and so on.

The complicated interaction between growth, precaution and the underlying societal values that weigh the two against each other, brings together three key themes in the global governance of machine learning. This is familiar territory for Europe: the dilemmas posed by ML and other new technologies have to be seen in the context of decades that Europe has spent grappling with the appropriate balance to be struck between seeing innovation as a source of growth or seeing it as a reason for caution. In the early years of this century caution was prioritized and the so-called precautionary principle became "a general principle of EU law" (Craig, 2012). There have been repeated subsequent efforts to balance precaution with innovation, but EU policy statements in this area now routinely begin by acknowledging the existence of a sizeable innovation gap between the EU and its international peers.[21] For example, in its 2018 communication on AI the European Commission notes that: "Europe is behind in private investments in AI which totalled around EUR 2.4-3.2 billion in 2016, compared with EUR 6.5-9.7 billion in Asia and EUR 12.1-18.6 billion in North America." (European Commission, 2018a)

It has proved difficult for Europe to transition swiftly towards prioritizing innovation. As our emphasis on the role of values suggests, it is not simply a matter of decision-makers mobilizing a certain level of political (or financial) capital. Societal norms and values are a constraining factor, a "centre of gravity" that it is difficult to shift. The role that can be played by underlying values is acknowledged in some of the European Commission's analyses of the EU innovation ecosystem. For example, in a 2018 document on research and innovation, the Commission highlights various

[21] For a counter-argument on the vitality of innovation in the EU, see "Innovation: Europe's Most Hidden Treasure?" in (Kalff & Renda, 2019).

technical factors that inhibit breakthrough innovations in Europe, such as limited venture-capital provision, but it also draws attention to "a deep-rooted aversion to risk". (European Commission, 2018c) Another way of expressing this risk-aversion, drawing on the definition of risk used by the International Risk Governance Center, is that European societies are less willing than their international peers to support policy choices that create uncertainty with respect to things people value. (IRGC, 2017, p. 5) In the context of the huge disruptive potential that a technology like machine learning possesses, a reluctance to accept uncertainty is an important constraint on potential governance choices.

While underlying attitudes to risk are not immutable, nor are they likely to change quickly. This is reflected in the need for the repeated European initiatives mentioned above that have sought to provide a pro-innovation counterweight to more cautious instincts and regulations. Attempts to offset the precautionary principle with an innovation principle have gained traction, but have also drawn sharp criticism.[22] The more recent "responsible research and innovation" (RRI) initiative is more promising, but its translation into practice is more difficult. Unless and until such initiatives generate a change in underlying attitudes towards risk, then the EU's approach to the governance of disruptive technologies will either have to go with the relatively risk-averse grain, or else accept a potential public backlash if unexpected adverse impacts emerge later.

On the face of it, a starting point of "deep-rooted aversion to risk" is difficult to reconcile with aspirations to be a leader in global technology governance. The question we address in the remainder of this report is whether there are ways for the EU to harness its uneasy balance of caution and growth in a way that might be a driver of global influence over machine-learning governance. Our answer is that such opportunities exist, so long as a pragmatic and focused perspective is maintained as to what is feasible.

If there is a trade-off between precaution and growth then we would expect countries (or regions) enacting fewer precautionary rules to see greater innovation and growth.[23] In a world in which both economic activity and technological innovations are highly globalized, the default would be for technologies developed in "high-growth" countries to spread internationally and to shape the terms of debates about governance norms and standards. The global penetration of high-growth technologies would tend to increase, in effect "exporting" at least some high-growth/low-precaution governance norms to other countries. This is part of what we saw in chapter 4: the

[22] See for example: European Risk Forum http://www.riskforum.eu/innovation-principle.html; https://ec.europa.eu/epsc/file/strategic-note-14-towards-innovation-principle-endorsed-better-regulation_en; https://encompass-europe.com/comment/the-innovation-principle; https://corporateeurope.org/en/environment/2018/12/innovation-principle-trap; https://www.greens-efa.eu/en/article/news/the-innovation-principle-is-a-regulatory-trojan-horse-from-the-industry/

[23] Precaution is not the same as state intervention. A strongly interventionist state could seek to promote innovation and growth rather than to protect non-growth values. As discussed above, the nature and impact of intervention will depend on the prevailing values in a given political system.

default governance landscape is heavily influenced by the dominant technologies being exported by countries prioritizing innovation and growth over ethical precautions.

5.2. A "niche leadership" governance strategy

However, the influence enjoyed by dominant technology exporters is not total. Depending on the nature of the norms being diffused internationally in this way, higher-precaution countries might make targeted defensive governance moves, equivalent to import controls designed to restrict the circulation of dangerous goods. At a minimum, such controls would create (or restore) for high-precaution countries a degree of domestic "defensive influence" over the governance standards that apply in their own territories to the technologies in question. But it is also possible that these controls might themselves turn out to be a source of potential external influence (or "exports") for a high-precaution country, if other countries also value the protections offered by the controls.

The evolution of the EU's GDPR provides a concrete example of this kind of global interplay of influence and counter-influence. The GDPR was (among other things) a response to the rapid global spread of technologies deemed to undermine one of the EU's core values, namely the protection of individuals' personal data. It provides the EU with a strong degree of "defensive influence" by establishing data-protection rules with which all technologies affecting EU citizens must comply. Increasingly, however, the GDPR is being seen as a source of potential external influence for the EU over the governance norms that apply to these technologies globally as well as within the EU. See, for example, a European Commission paper published in July 2019 (entitled "Data protection rules as a trust-enabler in the EU and beyond") which points to a trend of "upward convergence" on data protection standards, and welcomes the fact that GDPR adequacy provisions are becoming an increasingly important global benchmark (European Commission, 2019c). This capacity of the EU to extend its regulatory reach into other jurisdictions is sometimes referred to as "the Brussels effect". (Bradford, 2020)

Table 8: Interplay of domestic and global influence over technology governance

Scope			
	Comprehensive	Super-power	Walled garden
	Selective	Niche leader	Red lines
		Global	Domestic
		Reach	

The 2 x 2 matrix above provides a stylized way of mapping this interplay of domestic and global influence over technology governance. Each of the four cells describes a strategy that a country

might use to exert governance influence, divided according to the geographical reach of the influence (domestic or external), and on the scope of the influence sought (comprehensive or targeted). In each case, the basis of the influence might be market-driven, values-driven, or a mixture of the two. The GDPR example would position the EU in the two cells on the bottom line of the matrix: first, setting governance "red lines" that apply within its own jurisdiction (largely to set values-driven criteria on market-leading technologies developed in the US—located in the top-left "super-power" cell), but then leveraging the governance standards embodied in the GDPR to act as a "niche leader" setting data-protection standards that other stakeholders might adopt as their benchmark. This might involve other jurisdictions implementing similar rules—this is actively incentivized by the GDPR's requirement that "adequate" standards be in place before data transfer agreements with the EU are permitted. But it might also involve multinational corporates opting to apply EU standards globally—see for example Microsoft's response to the GDPR. (Brill, 2018)

On its own, this already amounts to a significant degree of external influence on global governance related to an issue that plays across numerous emerging technologies. However, it is also worth asking whether a values-based niche leader could "close the loop" by translating its values influence into a source of sectoral competitive advantage that would fuel innovation and growth (and thereby build a new source of influence on that dimension), in line with the EU Industrial Package adopted in March 2019.[24] In chapter 3 we highlighted that different normative and governance challenges can arise in different sectors/domains in which commercial applications for machine learning are developing. In chapter 6, we will suggest that there are targeted opportunities for the EU to develop *both* normative and competitive influence in some of these sectors, precisely because of the reputation it has already built as a leader in aligning technological progress with societal protections. Before that, however, in the next chapter we will consider the opportunities and risks confronting key global stakeholders in the current technology-governance landscape.

5.3. Stakeholder opportunities and risks

The decisions that the EU must take about its approach to the global governance machine learning do not exist in isolation. They are influenced by, and in turn are an influence on, the decisions made by other global actors. In the table spread over the pages that follow, we map some of the opportunities and risks facing the most globally significant stakeholders. In chapter 4 we noted that the global governance landscape is being shaped to a significant extent by the EU, the US and China. For that reason, we now highlight the opportunities and risks faced by stakeholders in these three areas, focusing in each case on three sectors: government, industry

[24] See https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_418

and the academic/scientific community. A number of themes recur in this map: notably, the strong incumbency benefits that stakeholders in the US and China already enjoy, which serves as a barrier to EU stakeholders trying to extend their influence. This lagging position of the EU is compounded by continuing fragmentation on important dimensions (from language and culture to market structure and financing), as well as by the relative under-development of the innovation ecosystem. However, the table below also highlights a number of strengths that the EU enjoys. These include: its increasingly entrenched status as a global “super-regulator” (Chander et al., 2020) that is able to wield extra-territorial influence via the so-called “Brussels effect” (already mentioned in the previous section); its experience (with the GDPR) of leveraging this influence to promote fundamental rights as the starting point for the governance of digital technology; and the potential market-making scale of its large public sector.

Table 9: Stakeholder opportunities and risks

OPPORTUNITIES AND RISKS FOR KEY STAKEHOLDERS						
Stakeholder Geography and/or category		**Motivations, goals**	**Influence** "+" (very low) to "+++++" (very high)	**Emphasis** "A" (growth/innovation) "B" (ethics/caution)	**Domain influence** Influence across selected machine-learning domains	**Opportunities/risks** Factors that might increase or decrease current influence
US	**GOVERNMENT**	• Maintaining technological and economic leadership in AI/ML • National security (1): constraint on citizens' data protection rights • National security (2): preventing China from global leadership in sensitive areas (eg 5G infrastructure) • Provide support to own industry. • Growth and consumption instead of sustainability.	+++	Primarily A, but growing focus on B	• Autonomous vehicles (AV): actively supporting • Healthcare (HC): innovative in regulatory approach • Public administration (PA): mix of support, application and contestation	• Network effects and "winner takes all" dynamics mean US incumbency benefits are particularly strong • Dynamic and responsive financial services sector, particularly the venture capital ecosystem • The global scale of leading US firms means that non-US regulators and fiscal authorities have significant influence • Deglobalization is a risk: trade tensions, supply chains, global market structures, domestic industrial policies, etc. • Geopolitical tensions could exacerbate these dynamics
	INDUSTRY (1) • "disruptive" players such as	• Broad-based dominance, which typically involves aggressive moves into new areas	+++++	Strongly A, but under growing external pressure with respect to perceived deficits on B	• AV: actively investing (e.g., Waymo LLC	• Network effects and "winner takes all" dynamics create high barriers to entry • Exceptional financial resources

	Google/Alphabet, Facebook, Amazon, etc)	• Focus on leadership across whole ecosystems rather than in specific sectors			subsidiary of Alphabet Inc)	• Strong political support (maintained with extensive lobbying activities) • Highly desired destination for global talent pool • Successive scandals (data breaches, misinformation, etc) mean the chances of domestic US regulation have increased • Global scale means leading players are affected by extraterritorial reach (both *de jure* and *de facto*) of non-US regulators (such as data protection and competition authorities in the EU)
	INDUSTRY (2) • "traditional" tech players such as IBM, Microsoft, etc	• Introduction of AI/ML into core business- and consumer-facing information technology (IT) functions • Selected pushes into leveraging AI/ML to develop leadership position in new sectors –for example, IBM Watson in healthcare • Trade-off between the need for a global market by the digital business incumbents and the push towards internalization by the government.	++++	A	• HC: actively investing and providing solutions (e.g., IBM Watson)	• Ready access to uniquely deep and competitive capital markets • Nature of business models and customer relationships limit their ability to collect and monetize data to the same extent as disruptors • The flip side is that more "traditional" US multinationals are not as vulnerable as disruptors if there is a backlash against "big tech" excesses • Vulnerable to de-globalization and to the efforts of individual countries to develop and protect their own AI/ML ecosystems
	OTHERS • academia and scientific institutions	• Leadership for research on AI/ML, management and digital transformation of education (MOOCS) • Talent attraction and management	+++++	Primarily A, but also B	• AV: actively researching and developing • HC: actively researching with	• Top academic organizations and journals still US-based • Positive synergies between academic and commercial leadership—job opportunities for graduates

		• Global positioning in potential high-growth AI/ML countries (through the creation of official affiliate campuses)			relevant contributions from the social sciences	• Potential negative synergies—suggestions of conflict of interest where research is funded by technology companies • Endowment model of funding not yet available in other countries • Geopolitical tensions could potentially disrupt/distort patterns of international education
CHINA	**GOVERNMENT**	• Rapid progress towards global technological and economic leadership in AI/ML • Increased domestic value-add throughout the supply chain—reduced reliance on foreign components and intellectual property • Ensure that AI/ML is deployed in ways that bolster rather than undermine the role of the government	++ (this reflects global influence; much higher domestically)	A		• Central control allows for unambiguous targeting of resources and corporate activity towards stated goals • Few constraints on development and deployment of AI/ML technologies, including creation of and access to huge reservoirs of data • Geopolitical tensions emerging as a brake on external influence • Walled garden structure of the digital market limits disruption by external players, but also limits external/global role of innovative domestic players (particularly consumer-facing)
	INDUSTRY	• Innovation in pursuit of the government's goal of developing a global leadership position	+ (this reflects global influence; higher domestically, but constrained by central government control)	A		• Access to huge quantities of data—owing to the size of the population and the comparative freedom to create and collect data • Relationship between the state and leading tech companies—a source of protection/promotion, but also free to interfere

EUROPE	**GOVERNMENT** • understood here as the EU institutions	• Ensure that the deployment of AI/ML is consistent with protection of fundamental rights (particularly in relation to privacy and data protection) • Ensure that the deployment of AI/ML does not undermine the (broadly egalitarian) social model • Build the innovation ecosystem and remedy Europe's commercial under-performance relative to the US and China • Increasing focus on sustainability as a pre-condition innovation and growth • Leverage the EU's relative strength as a "super-regulator" with extraterritorial reach	++	Primarily B, but actively seeking to expand on A	• AV: actively supporting • HC: actively supporting and innovative in regulatory approach • PA: actively supporting	• Influence can be boosted by the "Brussels effect", whereby rules set in the EU develop global traction—in the technology field, the GDPR is the paradigmatic case • Adoption and pursuit of a unified strategy is potentially complicated by differences at different levels of EU multi-level governance, as well as by differing preferences/interests among member states • Being a global "supplier" of ethical regulation of digital technology is costly, with no guarantee that the influence it creates will lead to commercial returns for domestic players (setting the rules for global leaders from elsewhere rather than developing global leaders) • Europe's large public sector is a potential source of influence and growth—building ethical rules into AI/ML procurement criteria could create a significant market for such technologies.
	INDUSTRY	• Protect and extend the strength of legacy players in sectors that may be disrupted by AI/ML (such as the automotive sector, where Europe has many leading players including PSA, Renault, Volkswagen)	++	A	• AV: actively investing • HC: actively investing and providing solutions • PA: mainly system integrator and research within EU funding programs.	• Market fragmentation across the EU's member states (as well as linguistic, fiscal, financial, regulatory and normative differences) complicates the process of scaling up • Given the prevalence of network effects and "winner takes all" dynamics, it will be

						difficult for European companies to dislodge other incumbents • The strong focus of EU regulation on data protection rights hampers domestic players in building and exploiting datasets to the same extent as rivals in other jurisdictions • A relatively large public sector in Europe represents an opportunity for market-building, but this will be constrained if strategies and investments are focused on national rather than EU-wide activities • The financial ecosystem—notably venture capital—is evolving, but is still not comparable to that of the US • Increasing geopolitical/geoeconomic competition between the US and China risks squeezing EU players, but it may also create opportunities (for example for Nokia and Ericsson in 5G infrastructure)
	OTHERS • academia and scientific institutions	• Develop an AI/ML innovation ecosystem that will better leverage high-level scholars and research—particular strength in computer science and engineering (with a strong background in humanities)	++	A and B	• AV: actively researching and developing • HC: actively researching with relevant contributions from the social sciences. • PA: actively researching with relevant contributions	• Fragmentation a constraint again, hampering the mobility of talent and the ability of universities to attract talent. • Cultural and linguistic differences could be an opportunity, but at present act as a barrier to entry or a reason to study elsewhere (either more competitive EU members states or global players such as the US and China) • Relatively few European countries/universities competing academically at global level, with the rest

					from the social sciences.	focused on locally. In the Times Higher Education Ranking 2019 only the UK is represented in the top ten (with three); in the top 50 there are three German institutions and one each for Sweden, Belgium and France

6. Implications: developing a "niche leadership" strategy

In the previous chapter, we have suggested that one of the options open to Europe in the current global technology governance landscape is to adopt (or take further) an "ethics first" niche leadership role. This would leverage the relative strength (compared to other major global technology players) that the EU has already built up in this area. What might such an ethics first approach look like? Our analysis so far suggests that it would have two elements, transversal and sectoral.

- The transversal component would codify ethical norms that are seen as non-negotiable, across all domains. The GDPR is the obvious example here, setting out EU-wide rules that seek to ensure that all technologies deployed within the EU (or affecting EU citizens) are aligned with fundamental rights relating to data protection set out in the TEU.
- The sectoral component would be less uniform, and probably less static, than the transversal. Instead of ensuring that a core ethical value is applied across the board, this sectoral or domain-based governance would entail identifying and operationalizing a desired balance between the range of ethical values (and other considerations, including growth and innovation imperatives) that arise in specific technological domains.

Both of these approaches to securing influence over technology governance could ensure the kind of domestic (EU-wide) "red lines" illustrated in the bottom-right quadrant of the matrix in chapter 5. But both could also serve as a source of the "niche leadership" illustrated in the matrix's bottom-left quadrant: projecting the EU's governance influence into other countries or regions where, as outlined in section 5.3, there may be growing demand for frameworks to help anchor disruptive technologies in norms and values that are considered societally foundational.

6.1. Priority machine-learning domains

Looking at machine learning specifically, where might the sectoral focus of a European ethics-first strategy lie? Much more work is needed in order to develop the necessary evidence base here, particularly given the large number of domains in which machine learning plays an increasingly important role. However, if we look at the three machine-learning domains discussed in chapter 3, we can perhaps point to preliminary hypotheses. Our suggestion here is that among the three domains—autonomous vehicles, healthcare and public administration—Europe might be able to take a governance lead on the second and third. The leadership that Europe has already developed in the area of privacy and data protection is one reason for this, but so are the possible advantages in these areas that result from a decades-long tradition of balancing innovation with a precautionary approach to new technologies. Regarding possible leadership on

autonomous vehicles, it is beyond the scope of the paper to evaluate the chances in this domain, given the strong US and China industry leadership. Furthermore this is a domain where issues of privacy and ethics may be less prevalent than in the other domains.

Healthcare and public administration raise particularly pressing normative questions for governance systems and stakeholders to resolve. This is particularly true for public-sector stakeholders, who bear ultimate responsibility for the health and welfare both of their citizens and their societies as a whole. The domains of medicine and public administration bear very directly on these core governance concerns. Medicine is obviously fundamental to human health, while public administration does much to determine the healthy functioning of society as a whole, including the legitimacy of the relationship between governors and governed. Many of the most advanced developments in healthcare rely on sharing sensitive private data with third parties, which is or should be subject to informed consent by individuals, careful safety management by those third parties and rigorous scrutiny by regulators and trust providers. Public administration provides perhaps the clearest instance in which governance *of* technology becomes increasingly important because of the growing potential for governance *by* technology as algorithmic decision-making becomes more and more prevalent. In both domains, a leakage of private information or a breach in security can have far reaching consequences for individuals, with potential irreversible negative consequences.

For these reasons, in the cases of both medicine and public administration it is potentially highly risky to introduce emerging technologies the impacts of which are subject to a significant degree of uncertainty. Europe's precautionary tradition, often seen as a brake on innovation, might paradoxically be a spur to innovation in cases like this, where the consequences of things going wrong could be stark. If so, as well as enabling Europe to secure widespread "domestic" buy-in for the roll-out of new technologies, it might also create new external opportunities. This might be a matter of other countries adopting similar rules as the EU on, say, genomic medicine. But it might also create commercial opportunities, positioning the EU as a centre for excellence in a variety of "ethically high stakes" technologies. A "made in the EU" or "coded in the EU" designation could become a powerful indicator that goods or services relying on machine-learning have been designed in compliance with the highest of standards and with an explicit view to minimizing (or at least making transparent) any potential risks to individual health or to wider social systems.

Before looking at the potential costs and benefits for Europe of such an ethics-first approach towards increasing global influence, it is worth pausing to consider the question of what could improve the influence and performance of the EU on global governance. Is a European ethics-first strategy better pursued by the EU system operating as a cohesive unitary actor, or through the disaggregated efforts of individual member states?

On the one hand, an instrument like the GDPR highlights the power of a single set of rules that apply uniformly across the EU and that can be presented to the rest of the world as a benchmark to be replicated. Albeit only in the field of privacy and data protection, the GDPR goes a long way to answering the famous Kissinger question: "Who do I call if I want to speak to Europe?" No matter who you call in Europe, they will turn to the GDPR as the basis of their reply. On the other hand, however, a one-size-fits-all regulatory response may be counter-productive in areas where there isn't yet the same degree of clarity and unanimity as evolved around the question of data protection.

At a time of profound technological flux, there may therefore be advantages to a two-stage process. The first stage would leverage the diversity of the EU member states in a process of "ethical entrepreneurship", allowing for experimentation with different governance models and with different balances between various ethical and other principles (including innovation, growth, etc). If particularly successful approaches are identified during this first stage, they could then be adopted across the whole EU in a second stage, whether as regulations, directives or softer sets of guidelines. We further suggest that such domain-specific EU-wide frameworks be designed according to the principles of "planned adaptation" (IRGC, 2016), so that any governance systems put in place are future-proofed against the inevitable evolution of both the new technologies in question and our understanding of their impacts on society.

6.2. Minimizing costs and maximizing benefits

It is important to note that seeking opportunities for the kind of ethics-first "niche leadership" approach discussed above would not be an all-or-nothing thing. To return to the relationship between precaution and innovation discussed in chapter 5, it would remain a core policy challenge find the optimal balance between these two. The suggestion here is not that Europe should depart the "non-ethics" rest of the field, or stop pursuing innovation-led strategies. The point being made here is the relatively modest one that there are ethical principles and technological domains where the EU may be able to expand its values-led influence and perhaps its competitive position too.

Where Europe does decide to adopt an ethics-first approach to building its governance influence, as with any strategy there will be both benefits and costs. Key potential benefits have been mentioned already: greater influence over the global governance of important new technologies, as well as possible associated commercial/innovation opportunities. In addition, the domestic benefits of aligning potentially disruptive technologies with European societal norms and values should not be overlooked, particularly at a time when technology is frequently cited as one factor in a pattern of declining social capital that is being recorded in many countries.

On the other side of the ledger, what are the possible costs or downsides of aiming for ethics-first niche leadership? Here are two:

- First, it could be seen as being too defeatist on the broader sweep of machine-learning innovation, tacitly accepting that other (the US and China in particular) will sustain or increase the lead that they have already built up. At the extreme, this in turn could lead to a weakening of the machine-learning ecosystem in Europe, paradoxically making it more difficult to actually design and deliver in the domains that have been targeted for niche leadership.
- Second, an ethics-first strategy could possibly lead to a backlash against the technologies in question if it creates the impression that the risks they pose greatly outweigh the potential benefits. Analogies have been drawn with the evolution of the debate in Europe surrounding genetically modified organisms. The analogy is far from perfect—not least because there does not seem to be any groundswell of anti-machine learning sentiment among consumers. However, it is at least possible that a call for gold-standard governance of machine-learning in a domain like public administration might lead to the response that "if the stakes are as high as you are claiming, then perhaps we are better off sticking with traditional methods".

It follows that if the EU were to seek to become a global-governance leader in certain machine-learning domains, one of the key associated policy challenges would be to explore ways of maximizing the benefits while minimizing the costs. Some of the building blocks for doing this are already in place. For example, on the benefits side, the European Commission's recent 5G initiatives focused on a range of industry verticals may mark a significant departure from previous 'all-purpose' policy frameworks. (Euroepan Commision, 2019a) More specifically related to machine-learning, the High-Level Expert Group on Artificial Intelligence (AIHLEG) provides a multistakeholder forum that aims to clarify the key ethical questions posed by AI, and to suggest policy responses to them. It might be possible for an entity like the AIHLEG to supplement its whole-of-AI focus with a workstream dedicated to identifying specific machine-learning domains where Europe is particularly well-placed to take on a niche leadership role. Distinct working groups for each of these priority domains could then be established under the auspices of the overall group, with a view to producing a strategy for the domain that would cover: (i) the distinct ethical challenges and trade-offs in the domain; (ii) the optimal governance responses to introduce in the EU; (iii) the potential demand for similar governance frameworks elsewhere in the world; and (iv) the possible commercial opportunities for European goods and services in this area.

The outputs from these priority-domain working groups could then feedback up to the AIHLEG and from there on to the Commission and other EU institutions to be acted upon. The proliferation of guidelines and initiatives in the area of technology ethics at the moment suggests that, as with the GDPR, one source of EU differentiation and concrete influence will be a willingness to codify rules, apply penalties and so on. This is a turbulent period for politics in much of Europe and more

widely, but the public sector remains a crucial source of decision-making trust and legitimacy, particularly in ethically charged areas such as the machine-learning domains we are talking about here.

When it comes to minimizing the potential costs associated with an ethics-first approach, one priority will be to ensure that this doesn't crowd out other activity in the machine learning innovation ecosystem. Again, there are institutional building blocks in place here, which could be maintained and developed, such as the European Innovation Council, and the successive rounds of support contained in the Horizon 2020 and Horizon Europe research and innovation programmes. Innovation support under these kinds of programme would be targeted at the priority machine-learning domains, to ensure that efforts to boost European influence over their global governance go hand in hand with an effort to increase European commercial opportunities in these areas, both domestically and in export markets. This would also be encouraged by a process of regulatory harmonization designed to reduce barriers to intra-EU trade in these machine-learning domains—analogous to (or possibly a component of) the Single Market for Trustworthy AI that has been proposed by the AIHLEG. (AI HLEG, 2019)

6.3. Ethical technology governance in the EU

In this chapter, we have made the suggestion that if the EU's objective is to increase its influence over the global governance of (and using) machine learning then it should consider focusing on those machine-learning domains where Europe's strong ethical focus gives it an advantage. This niche-leadership strategy would focus on developing EU machine-learning rules in domains where normative concerns are particularly prominent. Of the three cases we have focused on in this report, this would mean focusing on healthcare and public administration.

To what extent is this sort of "ethical regulation" of machine-learning a realistic proposition? There are a number of different types of evidence that might support its viability: (i) the use of similar governance strategies in other technology areas; (ii) problematic use of the machine-learning technologies in question; and (iii) existing moves to apply ethical rules to machine learning in the EU. Let us look at each of these in turn.

Analogies with other areas of EU governance

As discussed elsewhere in this chapter, the "ethics first" approach is a strategy that the EU has already pursued in the technology domain. The GDPR sets down data-protection standards with which all organizations operating in the EU must comply. This directly influences the governance landscape not just for "domestic" European actors, but for all global actors wishing to access the European market. In addition, there is a further indirect external influence, as the GDPR becomes a benchmark against which other regulatory standards are implicitly or explicitly measured. The strategic importance of the GDPR's external impact was noted by the Commission when marking

its one-year anniversary: “It is essential for us to shape the global field for the development of the technological revolution and for its proper use in full respect of individual rights.” (European Commission, 2019b) This sentiment is also echoed in relation to artificial intelligence, where the Commission stresses the need to ensure that “the Union’s values and fundamental rights are at the forefront of the AI landscape.” (European Commission, 2018a) Preparatory work at a sectoral level is already starting, with the Commission’s AIHLEG set to consider how its recommendations for ethical AI would apply in practice in a number of sectors (including healthcare).

It is also worth noting that the idea of using compliance with European technology rules as a marker that ethical concerns have been assessed and addressed would extend a similar approach that already applies in numerous other sectors. The CE marking provides consumers with a declaration that products in affected sectors meet stringent health, safety and environmental requirements. (European Commission, n.d.) Might there be an equivalent to the CE mark that developers would be required to use in order to distribute machine-learning algorithms in the EU/EEA?

Current use of machine learning in ethically charged domains

One example where ethical concerns have already been raised is in relation to policing and criminal justice. In the UK, the Law Society has called for greater oversight and regulatory protections covering the use of algorithms in the criminal justice system, citing worries about privacy breaches and biased decision-making. (Veale, 2019) A growing number of UK police forces have been experimenting with “predictive policing” software, much of which is provided by large US-based technology companies, mirroring the sectoral dynamics in many other technology areas. (Kelion, 2019) An indication of the expanding role of these companies across Europe is provided by the reported three-fold increase between 2014 and 2017 in the European revenues of Palantir, one of the global leaders in supplying machine-learning solutions to public-administration customers.[25] (Chapman et al., 2017) In the absence of a clear and robust European governance framework, the de facto governance regime will be strongly influenced by the norms and values embodied in the products and services supplied by leading companies in this area. This was the pattern that applied in relation to privacy before the GDPR was introduced to establish clear boundaries drawn up in line with European ethical values, priorities and trade-offs.

It will be some time before the use of machine learning in highly specialised healthcare fields, such as ML-based software as a medical device or precision medicine, becomes widespread. So the same level of market penetration by non-European companies does not seem to have occurred in this area as in public administration. (For example, IBM recently announced plans for its Watson for Genomics precision oncology software to be used for the first time in a European

[25] For an example of domestic UK technology, see the use of the Cambridge-developed Harm Assessment Risk Tool (HART): https://www.cam.ac.uk/research/features/helping-police-make-custody-decisions-using-artificial-intelligence

hospital.[26]) But even if it is not yet widespread, activity in precision medicine is increasing. And as noted in chapter 3, more rapid use of machine-learning tools is being recorded in other healthcare fields, such as medical imaging and direct-to-consumer genomics, which potentially raise a range of ethical implications, including standards of privacy, transparency and accuracy.

[26] See: https://www.labiotech.eu/medical/ibm-watson-genomics-european-hospital/

7. Recommendations (for the EU)

In this report we have mapped out the complicated and fast-evolving landscape for the governance of artificial intelligence and machine learning (AI/ML). We have highlighted a number of overarching normative considerations that apply in this area, but we have stressed the need to drill down to consider the specific governance challenges that arise when AI/ML is deployed in specific contexts. Generic principles are necessary but not sufficient to the task of governing AI/ML.

We have suggested that the EU may be able to carve out an influential role for itself by building on its strengths in the area of values-based regulation. But we have made the point repeatedly that ambitions in that direction must acknowledge at least two important and related caveats. The first of these caveats relates to the geopolitical and geo-economic backdrop. Competition is fierce, and multilateral approaches to global challenges of all kinds are under pressure. The EU does not have the luxury of developing its approach to machine learning in isolation and at its own pace. Others are active and leading in this area, notably the US and China, whose AI/ML sectors are technologically and economically dominant.

The second caveat relates to a potential trade-off between adopting a strongly normative focus and driving technological development and economic growth. In other words, even if an “ethics first” strategy is a source of significant influence on global governance arrangements, there is no guarantee that this will be reflected in economic gains, in the growth of a more vibrant technology sector, or in more efficient and fair public services. This potential tension between ethics and economic performance touches on governance dilemmas with which the EU has been grappling for decades. Broadly speaking, this involves the relationship between innovation on the one hand, and caution on the other hand. In general terms, the US and China have tended to strongly prioritise innovation as a source of growth, competitiveness and power, whereas Europe has attached greater weight to anticipating the potential adverse impacts (societal, environmental, economic, etc) that can accompany innovation.

This relationship between innovation and caution is thrown into particularly sharp relief by a technology as powerful and wide-ranging as machine learning. The choices that the EU makes in this area are likely to have important ramifications, both for its approach to emerging technologies more generally, and to its position in the wider global governance landscape.

7.1. Overarching recommendations

The analysis we have undertaken in this report leads us to five broad conclusions, which we suggest should frame the EU’s approach to governance in this area.

1. Focus on concrete problems in specific domains

As chapter 4 attests, there has been a flurry of activity in recent years around the construction of broad sets of principles and frameworks for the governance of AI and machine learning. This macro-level work is necessary, but not sufficient. Without moving swiftly to consider how stated principles and values apply (and potentially clash) in different ways depending on the specific context, there is a risk that AI principles and values will serve as window dressing, without much traction when it comes to real-world problem-solving. It is for this reason that we moved from our consideration (in chapter 2) of a series of overarching AI/ML priorities to assess (in chapter 3) the more specific challenges and trade-offs that arise in three domains: autonomous vehicles, public administration and healthcare. It goes without saying that these three domains are far from exhaustive—the work of understanding how machine learning is used, and how it should be governed, will require engaging with stakeholders in numerous other domains.

There are positive signs as to the EU's recognition of this need to build its overarching principles into more granular actionable insights. One broad example is that fact that the High-Level Expert Group on AI is in the process of revising the pilot version of its "assessment list", which aims to provide organisations with a set of benchmarks that can be used to ensure that key steps have been taken in order to deliver "trustworthy" deployments of machine learning. (Futurium, n.d.) More fundamentally, the Commission white paper *On Artificial Intelligence* stresses the importance of domain-specific factors in the regulatory framework that it envisages. (European Commission, 2020a) The white paper's focus on risk-based regulation is to be welcomed (it is less frequently used in the EU than in the US, where it is the baseline approach). Also welcome is the the white paper's use of "two cumulative criteria" for determining which applications of AI require particular regulatory attention: the first of these involves identifying those sectors where "significant risks can be expected to occur"; the second involves identifying the specific activities within those sectors which are likely to lead to such significant risks". This is a potentially powerful framework for advancing the governance of AI/ML, although it should be noted that there are likely to be significant divergences over what counts as a significant risk (not to mention how such risks should be weighed against the potential benefits that deploying AI/ML might create).

2. Build on the EU's normative priorities and strengths

The EU has begun to carve out an important governance role for itself in relation to the ethical implications of digital technologies. It is increasingly frequently described in this context using words and phrases such as "super-regulator" or "regulatory superpower". We suggest in this report that this role can be extended into the area of AI/ML. In the context of our three illustrative sectors, we suggest that the EU's normative strength and credibility position it well to influence the global governance of public administration and healthcare, where ethical factors are particularly important. Extending this logic out to encompass the range of "high-risk" applications

envisaged in the Commission's white paper, our suggestion is that the EU could have particular influence (relative to its global peers) in sectors where the risks identified are derived from what the white paper describes as "fundamental rights".[27]

We have described this as an "ethical niche leadership" strategy for the EU. It is possible that such a strategy could dovetail with an innovation strategy and become an engine of AI/ML development, uptake and exports for the EU. This seems to be envisaged in the white paper, which seeks to marry innovation (the "ecosystem of excellence") with ethics (the "ecosystem of trust"). This kind of mutual reinforcement might be possible. Perhaps an EU designation—such as "made in the EU" or "coded in the EU"—could become a gold standard in the market for AI/ML products and services in sectors where fundamental rights are at stake. Or perhaps the EU could take an assertive approach with third countries, bundling together its principles and products when exporting. A hypothetical example here might be if "adequacy" decisions like those in the GDPR rested not just on third countries' compliance with core principles, but on their use of EU-developed technologies that are certified as embodying those principles.

However, as noted above it should not be assumed that EU innovation and EU ethics will reinforce each other. It is possible that they will remain decoupled from each other, or that there will be uncomfortable trade-offs between them. To illustrate this, it is worth repeating a point made earlier in this report: in the area of data protection the EU's role as a regulatory heavyweight has not led to it becoming an innovation heavyweight in the technologies being regulated. Arguably this amounts to the EU positioning itself as a provider of global public goods rather than of private goods and services, taking on the cost of developing the normative underpinnings for the governance of an emerging technology, without necessarily increasing its share of associated global economic activity. That represents a significant contribution to global governance, but there may be questions as to its economic sustainability. In the context of a fragmented global context, it also raises geopolitical and geoeconomic questions that go beyond the scope of this report, such as how to weigh normative influence against more traditional measures of power such as economic or even military resources.

3. Balance unity and diversity

One of the elements of EU actorness that we are using in the TRIGGER project is cohesion, and it has been implicit in much of this report that the influence and effectiveness of the EU in global governance is strengthened when the EU is able to speak and act in unity. However, the value of diversity and experimentation should not be neglected. There are potential costs entailed in building greater cohesion if it stifles regulatory innovation or if it leads to high-stakes commitments

[27] The white paper states that AI risk should be assessed "in particular from the viewpoint of protection of safety, consumer rights and fundamental rights" (European Commission, 2020a, p17).

to a common governance strategy which may be outpaced by technological or societal developments.

As chapter 4 illustrated, there are numerous AI/ML governance initiatives under way in individual member states. There are also different approaches to high-level trade-offs such as between innovation and precaution. These initiatives and approaches should be seen as a valuable test bed for AI/ML governance strategies. If particularly successful approaches are identified at member-state level, they could strengthen or steer the approach adopted at the EU-level and thereby influence the global level.

4. Emphasize responsible and controlled data governance

As seen across this report, data is at the core of successull machine learning for good governance. Responsible collection, access, processing, sharing, use of data is key to achieving goals such as those around privacy and ethics. Data quality is key to achieving goals around accuracy and reliability. The data strategy adopted by the EC in January 2020 indicates a priortity that Europe's own companies are the ones capitalizing on data generated within its borders" (European Commission, 2020b). Europe must retain control over its data. This strategy also refers to the concept of data sovereignty. To address the complexity of AI/ML implications as well as to influence positively the ability of Europe to play a leading role in digital technology, one needs to have distinctive strategies for data and for AI/ML. The importance of measures such as those that allow and encourage data sharing within EU borders but not with non-EU countries must be acknowledged. For example, GAIA-X joint project of the EC, France, Germany and several companies to create a federated data infrastructure as the cradle of an EU data ecosystem is part of this strategy. It is based on data sovereignty "in the sense of complete control over stored and processed data and also the independent decision on who is permitted to have access to it". [28] There is a trend to gradually redistribute the data from the cloud to the edge and the devices, calling for a specific approach to "embedded" AI, which could contribute to reallocating value, and avoiding value capture by US and Chinese tech giants in key sectors such as automotive, manufacturing, and healthcare.

5. Develop future-proof governance mechanisms

AI/ML is an incredibly fast-moving technological area, and it is likely to remain so given the advances that are being made and the way these advances are being embedded in an ever-widening range of domains. This means that governance approaches are also going to have to evolve over time, or else risk becoming obsolete as AI/ML evolves and its ultimate impact on societies and economies becomes clearer. We recommend that this process of evolution should be built in as a core feature of the EU approach to AI/ML governance, using principles such as

[28] Project GAIA-X, Federated data Infrastructure, Federal Ministry for Economic Affairs and Energy, Federal Ministry of Education and Research.

those of “planned adaptive regulation”. (IRGC, 2016) This involves stipulating from the outset that regulations (and/or other governance mechanisms) will be revisited and amended over time on the basis of monitoring and feedback processes. The EU has already moved in this direction in the area of data protection—Article 97 of the GDPR requires the commission to report on the evaluation and review of the GDPR “by 25 May 2020 and every four years thereafter”. A similar approach should be adopted in relation to AI/ML, recognising that this may involve a more complicated and onerous process than for data protection. If it is a priority to develop a domain-specific approach to the governance of AI/ML, then there may need to be domain-specific monitoring, feedback and adaptation processes to ensure that salient developments affecting specific deployments are captured and dealt with.

7.2. Concluding reflections

The relationship between technology and governance is tightly coupled. This is particularly true of AI/ML, where increases in the prevalence of governance *by* technology intensify the need for robust and sustainable governance *of* technology. The co-evolution of these two processes—governance of and by AI/ML technologies—is likely to have a significant impact on the character of economies and societies around the world in the years ahead. It matters, therefore, what kind of role the EU is able to play in shaping governance approaches and frameworks in this area. With this in mind, we conclude this report by considering the following three questions. The first relates to governance of technology: can the EU increase its influence on the global governance of AI/ML technologies? The second relates to governance by technology: can the EU use AI/ML technologies in ways that might lead to new governance arrangements? And the third relates to the implications of all this for the EU itself: what are the implications of trends in AI/ML governance for the actorness of the EU?

1. Can the EU influence the global governance of AI/ML technologies?

Here, the answer would seem to be a strong “yes”. The evidence from data protection policy suggests clearly that the EU can wield extraterritorial power with its regulatory instruments and its overall approach to the governance of digital technologies. There are at least two channels for this external influence in the case of GDPR, which could be expected to recur in the area of AI/ML technologies if the EU adopted an analogous approach. The first is the so-called Brussels effect, which Chander, Kaminski and McGeveran (2019) describe as “a de facto mechanism, when market actors conform their global products to European rules”. The second is the adequacy process, a “deliberate legal export strategy” (ibid) which requires third countries to sign up to EU rules in order to avoid constraints on doing business in the EU, such as restrictions on data transfers in the case of GDPR. An AI/ML equivalent to the GDPR adequacy process might, for

example, involve third countries having to sign up to the EU's definition of, and regulatory requirements for, high-risk applications such as facial recognition.

Examples of where the EU could influence the governance of technology worldwide may include mandating the use of privacy-preserving techniques for handling input data in ML-based systems, or re-establishing legal requirements of explainability in algorithmic decision-making in public administration (Olsen, Slosser & Wiesener, 2019), or regulating facial recognition technology more strictly (European Commission, 2020a).

2. Can the EU use AI/ML to shape new ways of governing?

Governance is ultimately about the rules, systems, actors and behaviours that shape how decisions are taken. If one of the key characteristics of machine learning algorithms is that they can be trained to take certain decisions, then it stands to reason that this has potential implications for patterns of governance. It is striking that the impact of machine-learning decision-making systems has been so much greater in the private sector than the public sector over the past decade or so. There are numerous potential reasons for this—from inertia and skills gaps to the role of ethics and the rule of law—but it is not difficult to envisage increasing levels of machine-learning "disruption" in the public sector as these technologies become more mature. If so, then the EU could be well placed to help shape the process. The relatively large and diverse network of public sectors that exists across the EU and its member states provides ample space for the kind of experimentation discussed in the recommendations section above. Moreover, the overall scale of the public sector means the EU has potential market-making capacities in this area: a public-sector procurement programme for machine-learning tools that meet EU-approved criteria (for example, to do with human oversight, non-discrimination and data standards) could provide a significant incentive for private-sector activity.

In addition, shaping new and effective governance regimes could unleash both supply and demand in important sectors. To illustrate with examples given above, a possible positive consequence of mandating greater use of privacy-preserving techniques could be that the EU sets a global gold-standard for sharing medical and health data, thereby helping to overcome certain obstacles to the development of precision medicine. Similarly, mandating explainability in algorithmic decision-making would make ML-based diagnostics and decisions for precision medicine more acceptable to patients and medical professionals.

3. What are the implications of AI/ML governance for EU actorness

The question of "actorness" is one of the key conceptual threads running through the TRIGGER project. It will be taken up in more detail in Deliverable 4.4, which looks at cross-cutting themes across the EU's policy towards digital technologies. Actorness is also a central focus of the four thematic deep dives in Work Package 7, one of which relates to digital technology. However, in

anticipation of those more detailed discussions of actorness and technology, there are some tentative conclusions we can draw here in relation to our analysis of artificial intelligence and machine learning.

It is important to be clear about the core attributes of actorness as it has been defined by the TRIGGER project (see Deliverable 3.1). In this report so far, we have at times focused on the European Union in a holistic sense, encompassing both the EU institutions and the member states. One of the key reasons for this, as outlined in Chapter 4, is that in what remains an inchoate global governance landscape one of the key considerations is the relative position of European actors in general compared to the (other?) two technology superpowers, China and the US. By contrast, the TRIGGER definition of "actorness" focuses specifically on the EU institutions as they relate both to the member states and to external actors.

There are seven dimensions in the TRIGGER actorness model. Three are internal: (authority, autonomy and cohesion). Three are external (recognition, attractiveness and opportunity/necessity to act). The final dimension applies both internally and externally: credibility and trust. The list that follows considers to what extent the EU can be said to have high or low degree of actorness on each of these dimensions in the domain of AI/ML.

Internal dimensions

- **Authority** refers to the legal competences that the EU has in a specific area, as laid out in the treaties or in issue-specific agreements. There is no explicit competence relating to AI/ML in the treaties, but there are at least two treaty provisions on which the EU could claim authority to act in this area: the functioning of the single market, and the protection of citizens' fundamental rights. The EU enjoys strong authority in the area of data protection, which extends to those aspects of AI/ML that are covered by the GDPR. More broadly, authority for AI/ML is shared between the EU institutions and the member states: as policy evolves the respective roles of the member states and the institutions are being carefully calibrated. For example, following publication of its White Paper on AI, the Commission "will propose to the member states a revision of the Coordinated Plan [on AI] to be adopted by end 2020"—here we see the kind of issue-specific agreement mentioned above, with the EU seeking the authorisation to move ahead. In summary, the EU's authority in this area is shared and not particularly strong, but the framework conditions exist for it to increase in the future.

- **Autonomy** refers to the EU's capacity (including resources) to set priorities and act independently of the member states. As with the EU's authority, its autonomy is also shared with the member states. The EU institutions cannot unilaterally dictate policy or priorities in the area of AI/ML, but they enjoy significant agenda-setting powers. For

example, the work of the High-Level Expert Group on AI has played an influential role in framing the debate, particularly in relation to questions of ethics and values that have recurred throughout this report. Similarly, the White Paper on AI has continued and extended this framing process, particularly in relation to the need to orient policy around the twin objectives of technological and economic growth, on the one hand, and fundamental rights on the other. The White Paper makes other potentially significant framing moves too, such as its suggestion of a risk-based approach to the regulation of AI/ML. Looking at the idea of EU autonomy more broadly, in February 2020 the Commission published a Communication entitled "A European strategy for data", which signals the goal of enhancing autonomy and sovereignty in data governance, sharing and access (European Commission, 2020b).

- **Cohesion** refers primarily to the level of consistency between the EU institutions and the member states. (However, it may also be affected by intra-EU differences between the various institutions.) In Chapter 4, we highlighted a wide range of AI/ML initiatives that are taking place at member state level as well as the EU level. As with authority and autonomy, this points again to weak or nascent cohesion rather than something stronger. So too does the fact that a "Coordinated Plan on AI" has been developed. There is strong cohesion on the underlying principles that motivate the EU approach to governance in this area—ie, growth/innovation and fundamental rights—but the proliferation of AI initiatives reveal different approaches to balancing and implementing those objectives. It is important to note that cohesion isn't necessarily a normative goal: as we noted in Chapter 6, the ability of different member states to experiment with different AI/ML policies could lead to better EU-wide governance outcomes. The richness of European diversity must also be preserved.

External dimensions

- **Recognition** is the first of the external dimensions and it is determined by international perceptions of the EU in a given governance domain. The usual criteria for assessing EU recognition involves things such as whether the EU is viewed as an important convening power or negotiating partner. Given that the global governance landscape for AI/ML remains underdeveloped and fragmented, it is difficult to gauge the level of EU recognition in this sense. In the absence of a clear governance landscape, perceived leadership in this domain is currently being determined on the basis of how AI/ML is being used to drive innovation and growth. The US and China are recognised as the clear leaders in this respect. One way of seeing the blueprint contained in the Commission's White Paper on AI is as an effort to narrow the EU's "recognition gap" with the US and China by: (i) galvanising EU efforts to boost AI/ML innovation and growth, and (ii) making

sure that governance debates include a strong ethics/values focus. As we have discussed in this report, the EU may be able to leverage the strong external recognition it already enjoys in relation to technology ethics and values, notably in the area of privacy and data protection, where principles developed in the EU have been hugely influential.

- **Attractiveness** is defined primarily in terms of external actors perception of the advantages of cooperating with the EU in a given policy area. This will refer to material advantages (such as increased economic opportunities), but it can also include the EU being seen as an exemplar of "best practice", to be emulated. As things stand, the EU would not rank highly for attractiveness in AI/ML. The leadership of the US and China on innovation and growth gives them higher levels of attractiveness in terms of potential material gain. The EU's focus on the implications of AI/ML for fundamental rights has the potential to be a source of attractiveness, but it remains in its early stages. The same is true of the EU's wider data strategy, which is highly relevant for AI/ML. Initiatives such as GAIA-X, which aim to ensure a safe and secure European data infrastructure, could be a strong future driver of attractiveness.

- **Opportunity/necessity to act** is the last of the external dimensions of actorness, and it refers to the existence of external conditions that create a window of opportunity for the EU to play an increased global governance role. An analogy from the data protection field might be external developments such as the Snowden revelations or the Cambridge Analytica scandal, which heightened privacy concerns globally at a time when the EU was developing its privacy protections with the GDPR. It is not clear that external circumstances are currently similarly conducive to greater EU actorness on AI/ML. There are numerous contentious areas where AI/ML governance deficits might cause problems that gives the EU greater scope to act. These include domains we have focused on in this report, such as healthcare and aspects of public administration. Facial recognition is another prominent example. However, the weak global macroeconomic outlook is likely to prompt some countries to focus heavily on innovation and growth, rather than to follow the EU in balancing growth and fundamental rights and values.

Cross-cutting

- The **credibility and trust** dimension refers to the overall reputation of, and level of support for, the EU in a policy domain, both within and outside the EU. Our view is that it is too early in the development of the AI/ML global governance landscape for the EU to have built a strong reputation for credibility and trust. One of the first benchmarks of credibility is likely to be how well the EU can deliver on the vision for AI/ML that has been

set out in the White Paper and various initiatives that have accompanied and preceded it. As we have repeatedly noted in this report, the EU may be helped in this regard by the credibility and trust that it has built up in the area of privacy and data protection.

This report concludes with a general positive statement about the possible role of Europe on *governing AI/ML technologies*, in order to be able to benefit from the opportunities presented by *governing by technologies*. It remains to be seen in more details on how this can be implemented in various sectors and in various scenarios.

References

23andMe. (2017). *The science behind 23andMe's Genetic Weight Report.* 23andMe. Retrieved from: https://mediacenter.23andme.com/company/about-us/

Aebi, M.F.,Tiago, M.M.&Burkhardt, C.(2016).*SPACE I –Council of Europe Annual Penal Statistics: Prisonpopulations. Survey 2015*. Strasbourg:Council of Europe

Afuah, A., Tucci, C. L., & Viscusi, G. (2018). *Creating and Capturing Value Through Crowdsourcing*. Oxford: Oxford University Press.

AI HLEG. (2019). *Ethics guidelines for trustworthy AI*. Eurpean Commission.

AI Now. (2018). Algorithmic Accountability Policy Toolkit. Retrieved September 10, 2019, from https://ainowinstitute.org/aap-toolkit.pdf

AI Now. (2019). AI Now - A research institute examining the social implications of artificial intelligence. Retrieved from https://ainowinstitute.org/

Aisoma. (2018, April 01). *Importance of unsupervised learning in data preprocessing*. Retrieved from https://www.aisoma.de/importance-of-unsupervised-learning-in-data-preprocessing/

AItaskforce. (2018). *White Paper on Artificial Intelligence at the service of citizens*. *Version 1.0*. The Task Force on Artificial Intelligence of the Agency for Digital Italy - ai.italia.it. Retrieved from https://ia.italia.it/assets/whitepaper.pdf

Anastasopoulos, L. J & Whitford, A. B. (2018). Machine Learning for Public Administration Research, With Application to Organizational Reputation. *Journal of Public Administration Research and Theory*, *29*, 491–510, https://doi.org/10.1093/jopart/muy060

Anguelov, D. (2019, February, 12). *Drago Anguelov (Waymo) - MIT Self-Driving Cars* [Video file] Retrieved from https://www.youtube.com/watch?v=Q0nGo2-y0xY&feature=youtu.be

Angwin, J., Larson, J., Mattu, S & Kirchner, L. (2016, May). Machine Bias. *ProPublica.* Retrieved https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Annoni, A., Benczur, P., Bertoldi, P., Delipetrev, B., De Prato, G., Feijoo, C., … Junklewitz, H. (2018). *Artificial Intelligence: A European Perspective*. Luxembourg: Publications Office. https://doi.org/10.2760/936974

Autovista Group. (2019). *The state of autonomous legislation in Europe*. Retrieved August 20, 2019, from https://autovistagroup.com/news-and-insights/state-autonomous-legislation-europe

Bendett, S. (2018). How the Russian Military Is Organizing to Develop AI. *RealClearDefense* Retrieved from https://www.realcleardefense.com/2018/07/21/how_the_russian_military_is_organizing_to_develop_ai_303290.html

Bendett, S. (2019). Russia Racing to Complete National AI Strategy by June 15. *Defense One.* Retrieved from https://www.defenseone.com/threats/2019/03/russia-racing-complete-national-ai-strategy-june-15/155563/

Bernstein, J., Elsayed-Ali, S., Kochi, E., & Patel, M. (2018). *How to Prevent Discriminatory Outcomes in Machine Learning* [White Paper]. *Global Future Council on Human Rights 2016-2018*. World Economic Forum.

Bojarski, M., Testa, D. D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., . . . Zieba, K. (2016). End to End Learning for Self-Driving Cars. *arXiv:1604.07316 [cs.CV]*

Booker, B. (2018, August 19). HUD Hits Facebook For Allowing Housing Discrimination. *NPR.* Retrieved 19 September 2018, from https://www.npr.org/2018/08/19/640002304/hud-hits-facebook-for-allowing-housing-discrimination

Brabham, D. C. (2013). *Crowdsourcing*. London. UK: The MIT Press Essential Knowledge series.

Bradford, A. (2020) *The Brussels Effect. How the European Union Rules the World*. Oxford: Oxford University Pres

Brill, J. (2018, May). *Microsoft's commitment to GDPR, privacy and putting customers in control of their own data*. Retrieved from https://blogs.microsoft.com/on-the-issues/2018/05/21/microsofts-commitment-to-gdpr-privacy-and-putting-customers-in-control-of-their-own-data/

Brookings. (2019, May 22). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. *Brookings.* Retrieved from brooking: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/#footnote-9

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., … Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Future of Humanity Institute.

Carpenter, J., (2015, July 6). Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/?noredirect=on&utm_term=.4fd746684ee0

Chander, A., Kaminski, E. M and McGeveran, W. (2019). Catalyzing Privacy Law. *U of Colorado Law Legal Studies Research Paper No. 19-25*. http://dx.doi.org/10.2139/ssrn.3433922

Chapman, L. & Turner, G. (2017, February 24). Peter Thiel's Palantir Spreads Its Tentacles Throughout Europe. *Bloomberg*. Retrieved from https://www.bloomberg.com/news/articles/2017-02-24/peter-thiel-s-palantir-spreads-its-tentacles-throughout-europe

Cihon, P. (2019). AI & Global Governance: Using International Standards as an Agile Tool for Governance. *United Nations University. Centre for Policy Research*. Retrieved from https://cpr.unu.edu/ai-international-standards.html

Ciresan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2010). Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition. *arXiv:1003.0358 [cs.NE].*

Clark, G. (2017). *Industrial Strategy - Building a Britain fit for the future*. *Cm9528*. London, UK: HM Government.

Clark, G. (2018). Government should lead AI certification: Finkel. *Government News.* Retrieved September 1, 2019, from https://www.governmentnews.com.au/government-should-lead-ai-certification-finkel/

Clark, G., Hancock, M., Hall, W., & Pesenti, J. (2019, August 20). AI Sector Deal. *GOV.UK.* Retrieved from https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal#industrial-strategy-at-a-glance

Corea, F. (2018). AI Knowledge Map: how to classify AI technologies. *Medium.* Retrieved October 2, 2019, Retrieved from https://medium.com/@Francesco_AI/ai-knowledge-map-how-to-classify-ai-technologies-6c073b969020

Craig, P. (2012). Precautionary Principle. In C. P. Editor (Ed.) *EU Administrative Law.* http://dx.doi.org/10.1093/acprof:oso/9780199568628.003.0021

Cress, M. (2019). Russian President Vladimir Putin offers insight into the future of AI. *Artificial Intelligence Mania.* Retrieved September 9, 2019, from http://artificialintelligencemania.com/2019/06/24/russia-outlines-national-ai-strategy/

CRUK. (2015). Cancer risk statistics *Cancer Research UK*. Retrieved from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/risk

Danish Government. (2019). *National Strategy for Artificial Intelligence*. Ministry of Finance and Ministry of Industry, Business and Financial Affairs.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). Proxy Non-Discrimination in Data-Driven Systems. *arxiv.org/abs/1707.08120*

Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., … Hajkowicz, S. (2019). Artificial Intelligence: Australia's Ethics Framework [A Discussion Paper]. Australia: Data61 CSIRO.

DCJS. (2012). *New York State COMPAS-Probation Risk and Need Assessment Study: Examining the Recidivism Scale's Effectiveness and Predictive Accuracy.* Retrieved from

http://www.northpointeinc.com/downloads/research/DCJS_OPCA_COMPAS_Probation_Validity.pdf

de Laat, P.B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?. *Philos. Technol*. 31, 525–541. https://doi.org/10.1007/s13347-017-0293-z

Demiaux, V., & Abdallah, Y. S. (2017). *How Can Humans Keep the Upper Hand? The ethical matters raised by algorithms and artificial intelligence*. *Report on the public debate led by the french data protection authority (CNIL) as part of the ethical discussion assignment set by the digital republic bill*. Commission Nationale de l'Informatique et des Libertés (CNIL).

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248-255. Miami, FL: IEEE.

Dheeba J., S. S. (2011). *A CAD System for Breast Cancer Diagnosis Using Modified Genetic Algorithm Optimized Artificial Neural Network.* Berlin: Heidelberg.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In *Multi Classifier Systems. MCS 2000.* LNCS.

Ding Y, S. J. (2018). A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain. *Radiology*, 456-464.

Donatelli, S. (2018). How Switzerland is becoming a hub for artificial intelligence [Blog post]. Retrieved August 21, 2019, from https://www.sitsi.com/how-switzerland-becoming-hub-artificial-intelligence

Doneda, D and Almeida, V. A. F. (2016, July-August). What Is Algorithm Governance?, *IEEE Internet Computing*, *vol. 20*, no. 4, 60-63. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7529042&tag=1

Doshi-Velez, F., & Kim, B. (2017, March 2). Towards A Rigorous Science of Interpretable Machine *arXiv:1702.08608v2 [stat.ML]*

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., … Wood, A. (2017). Accountability of AI Under the Law: The Role of Explanation. *arXiv:1711.01134 [cs.AI]*

DUBAI FDI. (2018). *Artificial Intelligence - Investment Opportunity Brief*. DubaiAdvantage.

Dutton, T., Barron, B., & Boskovic, G. (2018). *Building an AI World - Report on National and Regional AI Strategies*. CIFAR.

e-estonia. (2019). Estonia accelerates artificial intelligence development. Retrieved July 22, 2019, from https://e-estonia.com/estonia-accelerates-artificial-intelligence/

Elliott, A. (2019). How Australia can make AI work for our economy, and for our people. Retrieved September 1, 2019, from http://theconversation.com/how-australia-can-make-ai-work-for-our-economy-and-for-our-people-113744

Engels, F., Wentland, A., & Pfotenhauer, S. M. (2019). Testing future societies? Developing a framework for test beds and living labs as instruments of innovation governance. *Research Policy*, *48*(9), 103826. https://doi.org/10.1016/j.respol.2019.103826

Ensign, D., Friedler, S., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). Runaway Feedback Loops in Predictive Policing. *arXiv:1706.09847 [cs.CY]*

EPFL IRGC (2018). *The Governance of Decision-Making Algorithms.* Lausanne: EPFL International Risk Governance Center. https://doi.org/10.5075/epfl-irgc-261264

Etherington, D. (2019, July). Waymo has now driven 10 billion autonomous miles in simulation. Retrieved from https://techcrunch.com/2019/07/10/waymo-has-now-driven-10-billion-autonomous-miles-in-simulation/

European Commission (n.d.) CE marking. Retrieved from https://ec.europa.eu/growth/single-market/ce-marking_en

European Commission. (2012). Consolidated Version of the Treaty on the Functioning of the European Union. Official Journal of the European Union, C 326/49. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012E/TXT&from=EN

European Commission. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, 4.5.2016(L 119). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2016:119:FULL

European Commission (2018a, April 25). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - Artificial Intelligence for Europe {SWD(2018) 137 final}*. *COM(2018) 237 final*. Brussels, 25.4.2018: COM(2018) 237 final. Retrieved from https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF

European Commission (2018b, April 25). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - Communication on enabling the digital transformation of health and care in the Digital Single Market; empowering citizens and building a healthier society*. Brussels, 25.4.2018: COM(2018) 233 final. Retrieved from https://ec.europa.eu/digital-single-market/en/news/communication-enabling-digital-transformation-health-and-care-digital-single-market-empowering

European Commission (2018c, May 15). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - A renewed European Agenda for*

Research and Innovation -Europe's chance to shape its future. Brussels, 15.5.2018: COM(2018) 306 final. Retrieved from https://ec.europa.eu/commission/sites/beta-political/files/communication-europe-chance-shape-future_en.pdf

European Commission. (2018d). EU Member States sign up to cooperate on Artificial Intelligence. Retrieved May 21, 2019, from https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence

European Commission (2019a, May 16). *Europe advancing in 5G – new wave of projects launched to accelerate 5G take-up in vertical industries*. Retrieved from https://ec.europa.eu/digital-single-market/en/news/europe-advancing-5g-new-wave-projects-launched-accelerate-5g-take-vertical-industries

European Commission (2019b, July 21) *General Data Protection Regulation shows results, but work needs to continue*. [Press release]. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/IP_19_4449

European Commission. (2019c, July 24). *Communication from the Commission to the European Parliament and the Council - Data protection rules as a trust-enabler in the EU and beyond –taking stock.* Brussels, 24.7.2019: COM(2019) 374 final. Retrieved from https://ec.europa.eu/commission/sites/beta-political/files/communication_from_the_commission_to_the_european_parliament_and_the_council.pdf

European Commission. (2020a). White Paper *On Artificial Intelligence - A European approach to excellence and trust* [White paper]. COM(2020) 65 final. Retrieved from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission (2020b). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - European strategy for data*. Brussels, 19.2.2020: COM(2020) 66 final. Retrieved from https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

European Union Agency for Fundamental Rights. (n.d.) . EU Charter of Fundamental Rights. Citizens' rights. Article 41 - Right to good administration. Retrieved from https://fra.europa.eu/en/eu-charter/article/41-right-good-administration

FDA. (2018). Software as a Medical Device (SaMD). Retrieved from https://www.fda.gov/medical-devices/digital-health/software-medical-device-samd

FDA. (2020). Artificial Intelligence and Machine Learning in Software as a Medical Device. Retrieved from https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Ferguson, A. (2018, June). How data-driven policing threatens human freedom. (*The Economist,* Interviewer). Retrieved from https://www.economist.com/open-future/2018/06/04/how-data-driven-policing-threatens-human-freedom

Fischer, S.-C. (2018, February). Artificial Intelligence: China's High-Tech Ambitions. *CSS Analyses in Security Policy*, *220*

Fjeld, J., Hilligoss, H., Achten, N., Daniel, M. L., Feldman, J., Kagay, S., & Singh, A. (2019). Principled Artificial Intelligence - A Map of Ethical and Rights-Based Approaches. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-019-0055-y

Freedman, H. D., (2017, June 27). A Reality Check for IBM's AI Ambitions. MIT Technology Review. Retrieved 19 September 2018, from https://www.technologyreview.com/s/607965/a-reality-check-for-ibms-ai-ambitions/

Future of Life Institute. (2019a). AI Policy - Australia - Future of Life Institute. Retrieved September 1, 2019, from https://futureoflife.org/ai-policy-australia/

Future of Life Institute. (2019b). AI Policy - Russia. Retrieved September 8, 2019, from https://futureoflife.org/ai-policy-russia/

Future of Life Institute. (2019c). AI Policy - Saudi Arabia. Retrieved August 21, 2019, from https://futureoflife.org/ai-policy-saudi-arabia/

Future of Life Institute. (2019d). AI Policy – France. Retrieved July 23, 2019, from https://futureoflife.org/ai-policy-france/

Futurium (n.d.) *Pilot the Assessment List of the Ethics Guidelines for Trustworthy AI*. Retrieved from https://ec.europa.eu/futurium/en/ethics-guidelines-trustworthy-ai/pilot-assessment-list-ethics-guidelines-trustworthy-ai

German Commission: Safeguarding Digital Sovereignty of Europe is Ethical Responsibility. Retrieved from https://dataethics.eu/german-data-ethics-commission-safeguarding-digital-sovereignty-of-europe-is-ethical-responsibility/

German Federal Government. (2018, November). Artificial Intelligence Strategy. AI Made In Germany. The Federal Government's Artificial Intelligence (AI) strategy. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwj5oLX6n_TnAhWyxIUKHXFJCoIQFjAAegQIBBAB&url=https%3A%2F%2Fwww.ki-strategie-deutschland.de%2Fhome.html%3Ffile%3Dfiles%2Fdownloads%2FNationale_KI-Strategie_engl.pdf&usg=AOvVaw0i5Aiv5H2bfyTzZAIJuHQh

Goldsmith, L., Jackson, L., O'Connor, A., & Skirton, H. (2013). Direct-to-consumer genomic testing from the perspective of the health professional: a systematic review of the literature. *Journal of Community Genetics*, 169-180.

Goodfellow, I. J., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative Adversarial Networks. *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems – Vol. 2,* 2672-2680. Montreal, Canada: MIT Press

Google. (2019). Artificial Intelligence at Google: Our Principles. Retrieved September 10, 2019, from https://ai.google/principles/

Government of Canada. (2019, May 16). *Canada and France work with international community to support responsible use of artificial intelligence.* [News release] Retrieved from https://www.canada.ca/en/innovation-science-economic-development/news/2019/05/canada-and-france-work-with-international-community-to-support-responsible-use-of-artificial-intelligence.html

Grzywaczewski, A. (2017, October). *Training AI for Self-Driving Vehicles: the Challenge of Scale*. Retrieved from Nvidia Developer: https://devblogs.nvidia.com/training-self-driving-vehicles-challenge-scale/

Gupte, R. (2019, May 8). AI in Cancer Detection: Are We There Yet? *Clinical Lab Manager*. Retrieved from https://www.clinicallabmanager.com/technology/ai-in-cancer-detection-are-we-there-yet-252

Hall, W., & Jérôme, P. (2017). Growing the artificial intelligence industry in the UK. Retrieved August 20, 2019, from https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

Haskins, C. (2019). Academics Confirm Major Predictive Policing Algorithm is Fundamentally Flawed. *Motherboard.* Retrieved from: https://www.vice.com/en_us/article/xwbag4/academics-confirm-major-predictive-policing-algorithm-is-fundamentally-flawed

Howe, J. (2006). Crowdsourcing: A definition. Retrieved November 7, 2017, from http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html

Hutchison, H. C. (2017). Russia says it will ignore any UN ban of killer robots. Retrieved September 8, 2019, from http://www.businessinsider.com/russia-will-ignore-un-killer-robot-ban-2017-11

Insights Team (2019, February 11). How Machine Learning Is Crafting Precision Medicine. *Forbes.* Retrieved from https://www.forbes.com/sites/insights-intelai/2019/02/11/how-machine-learning-is-crafting-precision-medicine/#20e59bfe5941

IRGC. (2016). *Planning Adaptive Risk Regulation.* Lausanne: International Risk Governance Center (IRGC). https://doi.org/doi:10.5075/epfl-irgc-228058

IRGC.(2017). *Introduction to the IRGC Risk Governance Framework,* revised version. Lausanne: International Risk Governance Council (IRGC). https://doi.org/10.5075/epfl-irgc-233739

ITU. (2018). AI Repository. Retrieved from https://www.itu.int/en/ITU-T/AI/Pages/ai-repository.aspx

ITU/WHO. (2018). FG-AI4H. Retrieved September 10, 2019, from https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx

Jain, S., Gillham, J., Mckellar, D., Duong, H., Saxena, K., & Rietschoten, J. van. (2018). *The potential impact of AI in the Middle East*. PwC.

Kalff, D. & Renda, A. (2019). *Hidden Treasures Mapping. Europe's sources of competitive advantage in doing business.* Brussels: Centre for European Policy Studies (CEPS). Retrieved from https://www.ceps.eu/wp-content/uploads/2019/09/Hidden-Treasures-Book_WEB.pdf

Kalra, N., & Paddock, S. M. (2016). *Driving to Safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?* Rand Corporation.

Kelion, L. (2019, February 4) Crime prediction software 'adopted by 14 UK police forces'. *BBC News.* Retrieved from https://www.bbc.com/news/technology-47118229

Kelly, É. (2019). Israel sets out to become the next major artificial intelligence player. Retrieved August 21, 2019, from https://sciencebusiness.net/news/israel-sets-out-become-next-major-artificial-intelligence-player

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S. (2017). *Human Decisions and Machine Predictions*, NBER Working Paper Number 23180

Kooi, T., Litjens, G., Ginneken, B. v., Gubern-Mérida, A., Sánchez, C. I., Mann, R., . . . Karssemeijer, N. (2016). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis 35*, 303-312.

Krafcik, J. (2018, December 05). *Waymo One: The next step on our self-driving journey*. Retrieved from Medium: https://medium.com/waymo/waymo-one-the-next-step-on-our-self-driving-journey-6d0c075b0e9b

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (p. 1097-1105). Lake Tahoe, Nevada: ACM.

Kumar, A., Shukla, P., Sharan, A., Mahindru, T., Sarkar, A., Nayan, A., … Raskar, R. (2018). *National Strategy for Artificial Intelligence*. *Discussion Paper*. NITI Aayog.

Liew, L. (2019, August). *Optimizing Your Trading Strategy to 2 million in Profits! - What is Curve Fitting (Overfitting)*. Retrieved from Algotrading101: Optimizing Your Trading Strategy to 2 million in Profits! - What is Curve Fitting (Overfitting)

Liu, S., Tang, J., Zhang, Z., & Gaudiot, J.-L. (2017). CAAD: Computer Architecture for Autonomous Driving. *arXiv:1702.01894 [cs.AR]*

Malgieri, G., & Comandé, G. (2017). Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*, *7*(4), 243–265.

Marchant, E. G., Allenby, R. B., Herkert, R. J. (Eds.) (2011). *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*. Netherlands : Springer

Massat, M. B. (2018). A Promising future for AI in breast cancer screening. *Technology Trends, Applied Radiology*, 22-25.

Miller, H., Stirling, R., Chung, Y., Lokanathan, S., Martinho-Truswell, E., New, J., … Scrollini, F. (2019). *Government Artificial Intelligence Readiness Index 2019*. Oxford Insights.

Ministero delle Infrastrutture e dei Trasporti. (2018). Smart Road: Via libera in Gazzetta Ufficiale alle strade intelligenti. Retrieved August 20, 2019, from http://www.mit.gov.it/comunicazione/news/smart-road/smart-road-libera-gazzetta-ufficiale-alle-strade-intelligenti

Ministry of Economic Affairs and Employment. (2017). *Finland's Age of Artificial Intelligence: Turning Finland into a leading country in the application of artificial intelligence - Objective and recommendations for measures*. *47/2017*. Ministry of Economic Affairs and Employment.

Ministry of Economic Affairs and Employment. (2019). *Leading the way into the age of artificial intelligence - Final report of Finland's Artificial Intelligence Programme 2019*. *2019:41*. Ministry of Economic Affairs and Employment.

Ministry of Enterprise and Innovation. (2018). *National approach to artificial intelligence*. *N2018.36*. Government Offices of Sweden.

MIT Technology Review. (2015, August 11). Why IBM just bought billions of medical images for Watson to look at? Retrieved from: https://www.technologyreview.com/s/540141/why-ibm-just-bought-billions-of-medical-images-for-watson-to-look-at/

MITA-NEMA. (2019, September). Digital Imaging and Communications in Medicine. Retrieved from DICOM: https://www.dicomstandard.org/

Moltzau, A. (2019). Scandinavian AI Strategies 2019. Retrieved July 11, 2019, from https://becominghuman.ai/scandinavian-ai-strategies-2019-16ecec9f17dc

MSIP. (2016). *Mid-to Long-Term Master Plan in Preparation for the Intelligent Information Society - Managing the Fourth Industrial Revolutio*. *Minister of Science, ICT and Future Planning (MSIP)*. Government of the Republic of Korea Interdepartmental Exercise.

National Academy of Sciences.(2018). *The Frontiers of Machine Learning: 2017 Raymond and Beverly Sackler U.S.-U.K. Scientific Forum*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25021

Nature [Editorial]. (2019, August 21). Protect AI panel from interference. *Nature*. 415 Vol 572 22 August 2019 Retrieved from https://www.nature.com/articles/d41586-019-02491-x

New, J. & Castro, D. (2018, May). How policymakers can foster algorithmic accountability. *Center for Data Innovation*. Retrieved from https://www.datainnovation.org/2018/05/how-policymakers-can-foster-algorithmic-accountability/

Ng, A. Y. and Jordan, M. (2002). - On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 841-848

NoCamels Team (2019). State of AI: Israeli Artificial Intelligence-Based Companies See Major Growth. Retrieved August 21, 2019, from https://nocamels.com/2019/03/artificial-intelligence-israel-major-growth-snc/

Nvidia. (2019, August). *Nvidia Drive - Software*. Retrieved from https://developer.nvidia.com/drive/drive-software

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, New York: Crown Publishers.*

OECD. (2017a). *Highlights from the OECD Science, Technology and Industry Scoreboard 2017 - The Digital Transformation: Italy*. OECD.

OECD. (2017b). *Summary of the CDEP Technology Foresight Forum Economic and Social Implications of Artificial Intelligence*. *OECD Conference Centre, Paris, 17 November 2016 - JT03408833*. Paris, France: Directorate for Science, Technology and Innovation.

OECD. (2019a). *Artificial Intelligence in Society*. Paris, France: OECD Publishing.

OECD. (2019b). OECD Principles on AI. Retrieved June 26, 2019, from https://www.oecd.org/going-digital/ai/principles/

OECD. (2019c). *Recommendation of the Council on Artificial Intelligence*. *OECD Legal Instruments*. OECD/LEGAL/0449.

Officialblogunio. (2019, April). *A short introduction to accountability in machine-learning algorithms under the GDPR*. Retrieved from https://officialblogofunio.com/2019/04/29/a-short-introduction-to-accountability-in-machine-learning-algorithms-under-the-gdpr/

Olsen, P., H., Slosser, L., J., Wiesener, C. (2019, June 12). What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration. iCourts Working Paper Series No. 162, 2019; University of Copenhagen Faculty of Law Research Paper No. 2019-84. http://dx.doi.org/10.2139/ssrn.3402974

OPSI. (2019). *Hello, World: Artificial Intelligence and its use in the Public Sector*. *Draft primer for public servants on the uses and considerations for AI in supporting public sector innovation and transformation*. OECD Publishing.

Papadakis, Z. G., Karantanas, H. A., Tsiknakis, M., Tsatsakis, A., Spandidos, A. D., Marias, K. (2019, March 13). Deep learning opens new horizons in personalized medicine (Review). *Spandidos Publications.* https://doi.org/10.3892/br.2019.1199

Poullet, Y. (2018). Is the general data protection regulation the solution? *Computer Law & Security Review*, *34*(4), 773–778. https://doi.org/10.1016/j.clsr.2018.05.021

Prescient & Strategic Intelligence. (2019). Europe Fully Autonomous Car Market is Expected to Reach 191.6 Billion by 2030. Retrieved from https://www.globenewswire.com/news-release/2019/05/29/1853802/0/en/Europe-Fully-Autonomous-Car-Market-is-Expected-to-Reach-191-6-Billion-by-2030-P-S-Intelligence.html

Press, G. (2016, March). Cleaning Big Data: Most time-consuming, least enjoyable data science task, survey says. *Forbes.* Retrieved from: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#1c9a50a06f63

Purdy, M., & Daugherty, P. (2016). *Why artificial intelligence is the future of growth*. Accenture.

pwc. (2019). Artificial intelligence – Switzerland lags behind global competitors. Retrieved August 21, 2019, from https://www.pwc.ch/de/press-room/press-releases/pwc-mr-switzerland_artificial intelligence-2019-en.pdf

Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., . . . Ng, A. Y. (2018). MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. *arXiv:1712.06957 [physics.med-ph]*

Rao, V. M., Levin, D. C., Parker, L., Cavanaugh, B., Frangos, A. J., & Sunshine, J. H. (2010). How widely is computer-aided detection used in screening and diagnostic mammography? *Journal of the American College of Radiology*, *Vol. 7*, Issue 10, 802-805.

Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. Y. (2018). Artificial Intelligence & Human Rights: Opportunities & Risks. *Berkman Klein Center Research Publication No. 2018-6.*, *September*. http://dx.doi.org/10.2139/ssrn.3259344

Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (2017). *Official Journal of the European Union.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745&from=GA

Reilly, C., (2018, May 13). Facial-recognition software inaccurate in 98% of cases, report finds. *cnet*. Retrieved from https://www.cnet.com/news/facial-recognition-software-inaccurate-in-98-of-metropolitan-police-cases-reports/Saurwein, F., Just, N., & Latzer, M. (2015). *Governance of Algorithms: Options and Limitations* (SSRN Scholarly Paper No. ID 2710400). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2710400

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. Retrieved September 10, 2019, from https://ainowinstitute.org/aiareport2018.pdf

Renda, A. (2019). *Artificial Intelligence - Ethics, governance and policy challenges*. CEPS.

RIVM. (2019, September 20). *Population Screening for Breast Cancer*. Retrieved from: https://www.rivm.nl/bevolkingsonderzoek-borstkanker

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., … Sechopoulos, I. (2019, September). Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists, *JNCI: Journal of the National Cancer Institute*, *111*(9), 916–922, https://doi.org/10.1093/jnci/djy222

Rossi, T. (2019, August). Autonomous and ADAS test cars produce over 11 TB of data per day. *Tuxera.* Retrieved from: https://www.tuxera.com/blog/autonomous-and-adas-test-cars-produce-over-11-tb-of-data-per-day/

Russia's National AI Strategy Takes Shape (2019). *Cyber Security Intelligence.* Retrieved September 9, 2019, from https://www.cybersecurityintelligence.com/blog/russias-national-ai-strategy-takes-shape-4356.html

Salathé, M., Wiegand, T., Wenzel, M., & Kishnamurthy, R. (2018). *Focus Group on Artificial Intelligence for Health*. FG-AI4H.

Saran, S., Natarajan, N., & Srikumar, M. (2018). In Pursuit of Autonomy: AI and National Strategies. Observer Research Foundation (ORF).

Saunders, D. (2017, July 17). The Bias-Variance Tradeoff. Retrieved from: https://djsaunde.wordpress.com/2017/07/17/the-bias-variance-tradeoff/

ScreenPoint Medical. (2019, September 27). Transpara: The software solution that will change the way you read mammograms. Retrieved from: https://www.screenpoint-medical.com/transpara

Siemens-Healthineers. (2019). *T*he Standard Foundation for Imaging and Image Management. Retrieved from: https://www.siemens-healthineers.com/de/services/it-standards/dicom

Sikkut, S. (2019). *Eesti tehisintellekti kasutuselevõtu eksperdirühma aruanne*. Riigikantselei, Majandus-Ja Kommunikatsiooni-Ministerium.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., . . . Demis, T. G. (2017). Mastering the game of Go without human knowledge. *Nature 550*, 354-359.

Singer, D. (2018). Israel's Artificial Intelligence Landscape 2018. Retrieved August 21, 2019, from https://becominghuman.ai/israels-artificial-intelligence-landscape-2018-a7bbecf280aa?gi=695569bd961e

Smallman, M. (2019). Policies designed for drugs won't work for AI. *Nature*, *567*(7 March 2019), 7.

Smith, A. (2018, August 30). Franken-algorithms: the deadly consequences of unpredictable code. The Guardian. Retrieved from

https://www.theguardian.com/technology/2018/aug/29/coding-algorithms-frankenalgos-program-danger

Smith, M. (2018). Can we predict when and where a crime will take place? *BBC News.* Retrieved from: https://www.bbc.com/news/business-46017239

SOPHiA. (2019). Democratizing Data-driven Medicine. Retrieved from: https://www.sophiagenetics.com/home.html

Sridhar, V., Subramanian, S., Arteaga, D., & Sundararaman, S. (2018). Model Governance: Reducing the Anarchy of Production ML. *Usenix Annual Technical Conference.* Boston.

Stix, C. (2018). *The European AI Landscape - Workshop Report*. European Commission.

Strategic Council for AI Technology. (2017). *Artificial Intelligence Technology Strategy (Report of Strategic Council for AI Technology)*. *March 31, 2017*. Strategic Council for AI Technology.

Tang, J., Liu, R., Zhang, Y.-L., Liu, M.-Z., Hu, Y.-F., Shao, M.-J., . . . Zhang, W. (2017). Application of Machine-Learning Models to Predict Tacrolimus Stable Dose in Renal Transplant Recipients. *Nature*.

The Conference toward AI Network Society. (2017). *Draft AI R&D Guidelines for International Discussions*. *28 July 2017*. The Conference toward AI Network Society.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2018). *Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. *Version 2 - For Public Discussion*. IEEE.

The Senate of the United States. (2019). *Algorithmic Accountability Act of 2019*.

Towers-Clark, C. (2019). The Cutting-Edge Of AI Cancer Detection. *Forbes.* Retrieved from https://www.forbes.com/sites/charlestowersclark/2019/04/30/the-cutting-edge-of-ai-cancer-detection/#6d5e73187336

Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., … Lin, H. (2017). FairTest: Discovering Unwarranted Associations in Data-Driven Applications, 401–416. https://doi.org/10.1109/EuroSP.2017.29

Tranberg, P. (2019, June). German Commission: Safeguarding Digital Sovereignty of Europe is Ethical Responsibility. Retrieved from https://dataethics.eu/german-data-ethics-commission-safeguarding-digital-sovereignty-of-europe-is-ethical-responsibility/

White House, D. (2019). Executive Order on Maintaining American Leadership in Artificial Intelligence - Issued on: February 11, 2019. Retrieved June 26, 2019, from https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

Veale, M. (2019). *Algorithm in the Criminal Justice System.* London: The Law Society of England and Wales. Retrieved from https://www.lawsociety.org.uk/support-services/research-trends/algorithm-use-in-the-criminal-justice-system-report/

Veale, M., & Brass, I. (2019, April 20). Administration by Algorithm? Public Management meets Public Sector Machine Learning. https://doi.org/10.31235/osf.io/mwhnb

Villani, C. (2018). *For a Meaningful Artificial Intelligence Towards a French and European Strategy*. Mission assigned by the Prime Minister Édouard Philippe.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99.

Wacket, M., Escritt, T., & Davis, T. (2017). Germany adopts self-driving vehicles law. Retrieved from https://www.reuters.com/article/us-germany-autos-self-driving/germany-adopts-self-driving-vehicles-law-idUSKBN1881HY

Walsh, T., Levy, N., Bell, G., Elliott, A., Maclaurin, J., Mareels, I. M. Y., & Wood, F. M. (2019). *The effective and ethical development of artificial intelligence: An opportunity to improve our wellbeing*. Report for the Australian Council of Learned Academies. Retrieved from https://acola.org/hs4-artificial-intelligence-australia/

Wang, Z., Ren, W., & Qiu, Q. (2018). LaneNet: Real-Time Lane Detection Networks for Autonomous Driving. Retrieved from https://www.semanticscholar.org/paper/LaneNet%3A-Real-Time-Lane-Detection-Networks-for-Wang-Ren/52a8866dfd2bce6a1169eba1a47ad2008ef4eecd

Wernick, M. N. (2010). Machine Learning in Medical Imaging. *IEEE signal processing magazine, 27(4)*, 25-38.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., … Schwartz, O. (2018). AI Now Report 2018. Retrieved September 10, 2019, from https://ainowinstitute.org/AI_Now_2018_Report.pdf

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*. London, UK: Nuffield Foundation.

World Economic Forum. (2019). Artificial intelligence: improving man with machine. Retrieved September 10, 2019, from http://reports.weforum.org/digital-transformation/artificial-intelligence-improving-man-with-machine/

WSJ. (2017). How Robots May Make Radiologists' Jobs Easier, Not Redundant. Retrieved from: https://www.wsj.com/articles/how-robots-may-make-radiologists-jobs-easier-not-redundant-1511368729

Xu, J., Yang, P., Xue, S., Sharma, B., Sanchez-Martin, M., Wang, F., . . . Parikh, B. (2019). Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. *Journal of Human Genetics Vol 138, Issue 2*, 109-124.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *In Computer Vision - ECCV 2014*, 818-833). Springer.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. *arXiv:1611.03530 [cs.LG]*

INSTITUTION EURASIAN INSTITUTE
OF INTERNATIONAL RELATIONS

trigger

Trends In Global Governance and Europe's Role

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement nº 822735. This document reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.