Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe

FINAL REPORT (D5)

Written by:
Andrea Renda (project leader), Jane Arroyo, Rosanna Fanni, Moritz Laurer, Agnes Sipiczki, Timothy Yeung (CEPS)
George Maridis, Meena Fernandes, Gabor Endrodi, Simona Milio (ICF)
Vivien Devenyi, Stefan Georgiev, Ghislain de Pierrefeu (Wavestone)

Internal identification

Contract number: LC-01528103
VIGIE number: 2020-0644

EUROPEAN COMMISSION

Directorate-General for Communications Networks, Content and Technology
Directorate A — Artificial Intelligence and Digital Industry
Unit A1 – Robotics and Artificial Intelligence Innovation and Excellence
Contact: Martin.Ulbrich@*ec.europa.eu*

European Commission
B-1049 Brussels

Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe

FINAL REPORT (D5)

Directorate-General for Communications Networks, Content and Technology
2021 EN

EUROPE DIRECT is a service to help you find answers to your questions about the European Union

Freephone number (*):
00 800 6 7 8 9 10 11

(*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you)

LEGAL NOTICE

This document has been prepared for the European Commission. However, it reflects the views only of the authors, and the European Commission is not liable for any consequence stemming from the reuse of this publication. The Commission does not guarantee the accuracy of the data included in this study. More information on the European Union is available on the Internet (http://www.europa.eu).

PDF	ISBN 978-92-76-36220-3	doi: 10.2759/523404	KK-03-21-189-EN-N

Manuscript completed in April 2021

1st edition

The European Commission is not liable for any consequence stemming from the reuse of this publication.

Luxembourg: Publications Office of the European Union, 2021

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This Study, completed for the European Commission, DG CONNECT, supports the Impact Assessment of the European Commission's proposed Regulation on a European approach to Artificial Intelligence (AI). The study is divided into four main sections.

Section 1 of the Study is dedicated to a comprehensive overview of existing evidence and prospective assessments of the risks and harms generated by AI for fundamental rights as well as for safety/security.

Section 2 provides a comparative overview of emerging national experiences in developing strategies and regulatory frameworks in this domain, with specific emphasis on risk governance.

Section 3 contains a detailed analysis of the results of the public consultation on the European Commission White Paper on Artificial Intelligence. It includes the analysis of 18 free-text questions from the consultation on the White Paper (6,667 free-text responses); and the screening of 408 position papers submitted to the public consultation.

Section 4 is dedicated to an assessment of the compliance costs generated by the proposed Regulation on Artificial Intelligence, including both administrative burdens and substantive compliance costs. The cost estimation is built on time expenditures of activities induced by the new requirements under the proposed Regulation.

Below, we describe our main findings in each of the sections.

Risks and harms created by AI applications

While acknowledging the outstanding opportunities offered by AI applications in a variety of domains, this section focuses on identifying potential areas of concern, which would elicit regulatory intervention. In this respect, the study finds strong evidence that certain uses of AI systems can significantly impact all fundamental rights recognised in the Charter of Fundamental Rights of the European Union. Such risks can occur in a variety of contexts, including business-to-business, business-to-consumer and also government-to-citizen situations. We survey several cases that came under the scrutiny of the Court of Justice of the European Union, the UN Special Rapporteur on freedom of opinion and expression, the EU High-Level Expert Group on AI, many EU national courts and data protection agencies, as well as numerous scientists and civil society organisations.
We conclude that AI has the potential to impact fundamental rights both positively and negatively: in this respect, it is rather the use of AI, as well as the design and governance arrangements built around such use, that determine the emergence of specific risks. AI risks are therefore heavily dependent on the context and the specific use.

Recurring **impacts on fundamental rights** are already found today in the following areas:

- **Bias and discrimination** are perhaps the most widely documented risks for fundamental rights generated by the use of AI systems. Even when discrimination is unintended, it can have far-reaching discriminatory impacts on key aspects such as gender, racial, social and other characteristics of target groups.
- Use of AI systems in specific contexts can also lead to the potential erosion of **human agency and autonomy**. Misinformation campaigns in combination with elaborated

recommendation engines, filter bubbles and echo chambers are increasingly powered by sophisticated AI. This can trigger addiction and opinion manipulation.

- The fundamental right to **freedom of expression and information, and the right to free elections** can be severely impacted by AI systems. Deliberate discrimination either against or for certain content by filtering and content removal practices by social media platforms was widely criticised in relation to both liability and censorship issues. Moreover, the right to free elections can be undermined when information dissemination is manipulated on social media platforms. Prominent examples such as the 'fake news' controversy around the 2016 US presidential election and the Cambridge Analytica/Facebook scandal in the wake of the Brexit referendum show the far-reaching impacts of AI-powered technology on free information and the democratic process.
- AI systems also impact **data protection and the right to respect for private and family life**. The free flow of data has ended up clashing in many occasions with the GDPR. A contentious issue in this regard is related to the use of AI-powered biometric identification and facial recognition technologies. Especially so-called 'second wave' biometrics deploy more elaborate technologies and algorithms, collecting highly sensitive and personal data.
- AI and automated decision-making used in government can impact **good administration, access to justice and the right to a fair trial**. Areas such as predictive policing, law enforcement, risk modelling and social scoring were found to create privacy and data protection risks, introduce new biases, and/or intentionally discriminate against individuals, for example ignoring or discriminating those for which social support systems are designed in the first place. AI used for predictive policing and law enforcement are already operating in more than half of the EU member states, and they have been found to threaten the EU right to be free from interference due to potentially unlawful data collection and arbitrary risk scoring, as well as raising questions around the right to a fair trial and innocence of the defendant.
- Further, the **impact of AI on specific vulnerable groups, such as migrants,** has come under scrutiny in several instances.
- The study identifies **impacts of AI on other EU fundamental rights**, including consumer protection, the right to freedom of assembly and association, as well as sustainability and protection against sustained impairment of the living standards of future generations.

The study also looks at **risks for safety and security** through a systematic literature review. We find that the nature of AI systems, as well as the context in which AI is deployed, may pose threats to several aspects of safety and security. Importantly, the risks generated by AI can emerge at various phases of the product lifecycle, from the design and development phase of AI to the deployment and post-deployment phases. We also find evidence that AI product safety and liability risks are exacerbated by the inadequacy of current civil liability rules in addressing AI systems. Characteristics such as connectivity, opacity, data dependency and autonomy are particularly important in this respect. In specific contexts such as healthcare, safety stances become even more important given the critical nature of the systems to be operated with the support of AI. In many domains, the quality of datasets is key in training AI systems: any failure in the initial data may cause incorrect outcomes and function erroneously throughout its application period, invalidating the entire AI system. Ensuring that the data source is trustworthy and accurate is key to preventing safety issues caused by AI. Human bias within the training dataset is a common issue in automated systems and can also compromise the safety of an AI system.

International experience of AI policy: emerging policy frameworks

The study analyses a number of selected international experiences on designing and implementing policy frameworks for the responsible use of AI, with a specific focus on the

governance of AI risks. The landscape appears already very heterogeneous and in constant evolution: however, no country to date has attempted to introduce a comprehensive, horizontal regulatory framework for Artificial Intelligence. Among the observed experiences:

Australia's voluntary framework is characterised by eight voluntary AI ethics principles in addition to specific business guidance for implementation. A recent public consultation emphasised the principles of fairness, transparency, contestability and accountability as important, while the development of AI standards is likewise a national priority.

Canada, with its Directive on Automated Decision-Making, was among the first countries to introduce a regulatory framework in 2018. The Directive on Automated Decision-Making establishes six elaborated risk classification levels which in turn result in different impact level requirements.

In **Germany**, the Data Ethics Commission proposed a risk classification that establishes a sector-neutral five-level scale of 'criticality' on which AI systems are classified according to the degree of potential harm, proposing a full or partial ban on systems categorised in the fifth level due to their untenable potential for harm.

Japan's Contract Guidelines on Utilisation of AI and Data are centred around the main challenges and unresolved issues of model contract clauses and additional factors to be considered in the preparation of contract clauses for concluding contracts on data or AI software/technology.

Singapore's model governance framework on AI is a voluntary framework and by design algorithm-neutral, technology-neutral, sector-neutral, scale and business-model-neutral, and is accompanied by an Implementation and Self-Assessment Guide for Organisations. The key areas covered by the model framework are the set-up of internal governance structures and measures for businesses, determining the level of human involvement in AI-augmented decision-making, and considering the risk on the probability and severity of potential harm caused by the AI system, as well as assessing operation, maintenance and data management, stakeholder interaction and communication, and opt-out principles for consumers.

The **United Kingdom** issued a guide on using AI in the public sector to support its administration to meet user needs, and to implement AI ethically, fairly and safely, including a guidance on choosing the most appropriate machine learning technique for a certain administrative process. The UK's Information Commissioner's Office Guidance on AI Auditing, in particular, is a voluntary framework for organisations covering best practices for assessing data protection risks associated with the use of AI systems, implementing accountability and governance mechanisms; ensuring compliance with the substantive data protection requirements such as lawfulness, fairness, and transparency; practical measures to aid security and data minimisation; and ensuring the protection of individual rights in AI systems.

The **United States** issued a draft Guidance for Regulation of Artificial Intelligence Applications, supposed to inform the development of regulatory and non-regulatory approaches by US agencies that interact with AI systems, thus establishing a sectoral approach. The draft guidance also seeks to reduce barriers to the use of AI technologies while upholding civil liberties. The document includes details on the definition of AI, risk assessment and management approaches, ways to avoid prescriptive regulation, and establishing voluntary conformity assessment standards. At the state level, the New Jersey proposed Algorithmic Accountability Act goes a step further by mandating certain entities to carry out impact assessments on high-risk AI. The Automated Decision System Impact Assessment would include several cost-benefit analysis on data, security and risk aspects.

Several other relevant frameworks have been introduced at the international, minilateral or multilateral level, as well as by the private sector. They include frameworks developed in the Nordic-Baltic region, UNESCO, the OECD, the G20, the International Telecommunications Union, and the Global Partnership on AI in the context of the G7. These entities, together with non-state actors and several multi-stakeholder initiatives, shape the global AI governance landscape. ISO/IEC and the IEEE Standard Association as two important international standards bodies also contribute to these developments by creating AI-specific standards. Research institutions such as AI Now further shape the conceptual work on AI oversight, accountability and auditing, for instance with the Algorithmic Impact Assessment to evaluate the use of AI systems in public agencies.

Summary of the results of the open public consultation

Section 3 of the study provides analyses of stakeholders' responses to the open public consultation on the European Commission White Paper on AI (2020).

In the position papers, the most important point for many respondents was the **definition of 'high-risk'**. Many respondents believe that the definition of high-risk in the white paper is unclear or needs improvement (at least 18% of all position papers, 74 out of 408). Many find that the binary classification in high vs. low is too simplified and some propose introducing more levels of risk. Some think that the definition is too broad, while others believe that it is too narrow.

Another important issue in the position papers was the proposed **voluntary labelling scheme** (at least 52). At least 21 position papers are sceptical of labelling, either because they believe that it will impose regulatory burdens (especially for SMEs) or because they are sceptical about its effectiveness. At the same time, at least eight position papers are explicitly in favour. Stakeholders also address a variety of other issues (see chapter 3)

Regarding the **definition of AI**, around 15.7% of stakeholders mention that they disagree with the definition of AI by the High Level Expert Group on AI and the Commission. At least 9.3% state that the definition is too broad (37), some of these stakeholders highlight that a broad definition risks over-regulation and legal uncertainty, and is not specific enough to AI. At the same time, at least 6.6% believe that the definition is too narrow (27) and can miss important dimensions of AI.

Regarding expected **costs,** up to 84% of position papers do not explicitly mention concerns about costs that could be imposed by a regulation on AI (344). At the same time, at least 11% of position papers (46) mention compliance costs in general, and at least 7% of position papers (29) (also) mention administrative burdens in particular.

Around 23% of position papers address potential **institutional structures for governing AI** in the EU. 10% of them are in favour of a new EU-level institution, with at least 6% favouring some form of a new EU AI agency (24) and at least 4% a less formalised EU committee/board (15). At the same time, at least 3% of position papers are against establishing a new institution (14) and several mention other types of arrangements.

More than half of the position papers do not mention the proposed **regulatory requirements for high-risk AI** from the white paper (human oversight, training data, data and record-keeping, information provision, robustness and accuracy). Many position papers (at least 23%), however, generally agree with the White Paper's approach to regulatory requirements

for high-risk AI. At least 12% generally disagree, and several stakeholders express other opinions (12%).

Based on the free-text responses to the open-ended questions in the open public consultation questionnaire, some key findings are:

- Regarding **other actions that should be considered based on the white paper** (open question - OQ 1), several respondents emphasize the importance of Skill building (58 respondents); the facilitation of data sharing & access (21 respondents); and the importance of a global approach to AI and international cooperation or standards (20 respondents).
- Regarding **other actions to strengthen the research and innovation community** (OQ3), some free-text responses refer to the network of existing AI research excellence centres (39) and are sceptical of lighthouse research centres (22).
- When asked about **other tasks for specialised digital innovation hubs** (OQ4), some stakeholders emphasize the importance of support for partnerships between SMEs, large enterprises and academia (33) and knowledge transfers to support the development of AI expertise for SMEs (27).
- When asked about other opinions on **whether concerns can be addressed by applicable EU legislation** (OQ6), several respondents write that current legislation has gaps or new legislation is required (47), while others think that more research is needed to express an opinion (39) or caution against overregulation (24).
- When asked whether "the **introduction of new compulsory requirements should be limited to high-risk applications**" (OQ7), respondents note that the definition of 'high risk' is unclear and more details are needed (33) and that the binary distinction between high/low risk is too simplified (24).
- Regarding the "**AI application or use that is most concerning ('high-risk')** from your perspective" (OQ9), stakeholders mentioned applications related to autonomous weapons (41), biometric identification (34), applications in critical infrastructure (e.g. electricity, water supply, nuclear) (28) and others.
- Regarding **other enforcement systems for AI rules** (OQ12), some stakeholders are in favour of (ex-ante) fundamental rights assessments (19, partly coordinated response by NGOs), while others are in favour of self-assessments (14) or independent external bodies/experts to ensure assessments (11).
- Regarding "**any further suggestions on the assessment of compliance**" (OQ13), some respondents are in favour of independent external assessments (32, partly coordinated response by NGOs, but also car makers), and other want to avoid overly burdensome procedures (29)
- When it comes to "**any further considerations regarding risk assessment procedures**" (OQ15), some stakeholders mention that risk assessments need to be repeated in case of changes after placement on the market (16) and that clearer definitions of what constitutes e.g. "important changes" are needed (13).

Assessment of the compliance costs generated by the proposed regulation on AI

Section 4 contains an assessment of compliance costs generated by the proposed Regulation on AI. The cost estimation is based on consolidated methodologies for the assessment of administrative burdens and substantive compliance costs. This study in particular employs the cost model developed by the Federal Statistical Office of the German government, which has the additional advantage of featuring standardised tables with time estimates per administrative activity and level of complexity.

The assessment considers the five regulatory requirements listed in the AI White Paper, and computes an estimate of cost for each requirement by identifying necessary activities and assigning to each activity an estimate of time needed to complete the task. Based on standardised cost tables and estimated levels of difficulty, we identified the main activities involved to fulfil each requirement. To verify our estimates, the methodology and the results of the cost assessment were presented in two workshops involving several experts and industry stakeholders; we then followed up with several bilateral discussions with stakeholders in order to ensure the representativeness of the figures selected. Similarly, accreditation bodies and standardisation organisations were invited to another workshop to discuss the team's estimates on the costs of conformity assessment procedures. We also performed extensive benchmarking with other similar procedures (e.g. in the domain of product safety, or medical devices) to check the accuracy of our estimates.

The estimated compliance cost of each requirement for one "AI unit" (assuming EUR 170,000 of average development costs) is summarized below:

- Training data: EUR 2,763
- Document and record keeping: EUR 4,390
- Information provision: EUR 3,627
- Human oversight: EUR 7,764
- Robustness and accuracy: EUR 10,733

The estimated annual labour compliance cost for a single AI product is EUR 10,977. Together with the purchase of external data and services, as well as hiring additional staff, this cost rises to **EUR 29,277**. We then performed an assessment of the activities that would take place also without an additional regulatory intervention (so-called "business as usual", or BAU factor) as well as learning effects over time, and concluded that **cost estimates for 2025 would be reduced by around 36%.**

The total compliance cost was then extrapolated to the estimated market size and projected to the future. We find that **the total compliance cost for the global AI industry is estimated to range from EUR 1.6 billion to EUR 3.3 billion in 2025**, assuming that only 10% of the AI units will be subject to the regulatory requirements (i.e. those identified as "high risk").

The second part of the cost assessment concerns the costs spent on a certification process of a regulated AI product through a conformity assessment. First, we estimate the costs of a conformity assessment of a single product (an "AI unit") under the EU-type examination procedure, relying on both a bottom-up approach and benchmarking to reach final cost estimates. Second, we estimate the cost of a conformity assessment of an AI unit under a Quality Management System (QMS). Under this procedure, the cost also includes the one-off expense of setting up a QMS (including benchmarking and validation from different experts). We conclude that under reasonable assumptions, **obtaining certification for an AI unit through the EU-type examination may cost on average EUR 16,800-23,000, roughly 10-14% of the development cost.** On the other hand, **setting up a new QMS may cost EUR 193,000-330,000 upfront plus EUR 71,400 yearly maintenance cost**. Most of the costs could be shared among different AI products in case the entity in question developed more than one AI product.

The cost assessment is, as inevitable, only a tentative exercise given the uncertainty related to the final content of the regulation, the portion of AI products that will be qualified as high risks, the complexity of the AI industry and value chains, the difficulty of predicting whether businesses will prevalently rely on pre-trained AI systems or develop and deploy AI in house, among other issues. More specifically:

- An important challenge is the **complexity of the AI ecosystem**. Product developers may purchase another AI system and embed it into another product. Compliance cost will thus vary depending on whether the purchased system has obtained certification and also whether the development of the new product involves additional data inputs and training. The complex ecosystem would potentially involve a complex sharing of liabilities.
- **Sufficiently equipped and qualified notified bodies are in extremely short supply**. They generally do not perform type examination of products containing software under the Medical Devices Reguation (MDR). Exhaustively testing a software is considered to be impossible. Besides, it is unclear whether the AI regulation would require actual auditing of training data and record-keeping, which may imply additional or regular audits as new training data may keep flowing in.
- Without defining clearly the requirements and testing procedures, notified bodies find it **difficult to come up with a cost estimate of the conformity assessment**. Some one-off costs, such as staff training, legal fees and any machinery and equipment needed, may fence off small notified bodies and thus the cost of the conformity assessment may be much higher in the early years of the regulation.
- We made the methodological choice to **exclude the costs of external legal advice and consultancy fees from the cost estimates**. This choice is grounded in the observation that these cost items are largely influenced by (i) the size of a company and the availability of in-house expertise, (ii) the preference of each company, and (iii) the complexity and stringency of the regulatory requirements in the proposed regulation.

RÉSUMÉ EXÉCUTIF

Cette étude, réalisée pour la Commission européenne, DG CONNECT, vient appuyer l'étude d'impact de la proposition de Régulation de la Commission européenne concernant une approche européenne de l'intelligence artificielle (IA). L'étude est divisée en quatre sections principales.

La **section 1** du rapport est consacrée à une vue d'ensemble des preuves existantes et des évaluations prospectives des risques et des préjudices générés par l'IA pour les droits fondamentaux ainsi que pour la sûreté/sécurité.

La **section 2** fournit un aperçu comparatif des expériences nationales émergentes en matière de développement de stratégies et de cadres réglementaires dans ce domaine, avec un accent particulier sur la gouvernance des risques.

La **section 3** contient une analyse détaillée des résultats de la consultation publique sur le livre blanc de la Commission européenne sur l'Intelligence Artificielle. Elle comprend l'analyse de 18 questions à texte libre de la consultation sur le livre blanc (6 667 réponses à texte libre) et l'examen de 408 prises de position soumises à la consultation publique.

La **section 4** est consacrée à une évaluation des coûts de mise en conformité générés par la proposition de réglementation sur l'intelligence artificielle, incluant les charges administratives et les coûts substantiels de mise en conformité. L'estimation des coûts est basée sur le temps consommé pour assurer les activités induites par les nouvelles exigences de la réglementation proposée.

Nous décrivons ci-dessous nos principales conclusions dans chacune des sections.

Risques et dommages créés par les applications de l'IA

Tout en reconnaissant les opportunités exceptionnelles offertes par les applications de l'IA dans de nombreux domaines, cette section se concentre sur l'identification des sujets de préoccupation potentiels, qui induiraient une intervention réglementaire. À cet égard, l'étude apporte des preuves solides sur le fait que certaines utilisations des systèmes d'IA peuvent avoir un impact significatif sur tous les droits fondamentaux reconnus dans la Charte des droits fondamentaux de l'Union européenne. De tels risques peuvent survenir dans divers types de relations, que ces dernières soient d'entreprise à entreprise, d'entreprise à consommateur ou encore de gouvernement à citoyen. Nous étudions plusieurs affaires qui ont été examinées par la Cour de justice de l'Union européenne, le rapporteur spécial des Nations unies sur la liberté d'opinion et d'expression, le groupe d'experts de haut niveau de l'UE sur l'IA, de nombreux tribunaux nationaux de l'UE et des agences de protection des données, ainsi que de nombreux scientifiques et organisations de la société civile.

Nous concluons que l'IA pourrait impacter les droits fondamentaux, tant positivement que négativement : à cet égard, c'est plutôt l'utilisation de l'IA, ainsi que la conception et les dispositions de gouvernance construites autour de cette utilisation, qui conduisent à l'émergence de risques spécifiques. Les risques liés à l'IA dépendent donc fortement du contexte et de l'utilisation qui en est faite.

Des **impacts récurrents sur les droits fondamentaux** sont déjà constatés aujourd'hui dans les domaines suivants :

- Les biais et la **discrimination** générés par l'utilisation des systèmes d'IA sont peut-être les risques pour les droits fondamentaux les plus largement documentés. Même lorsque la discrimination n'est pas intentionnelle, elle peut avoir des impacts discriminatoires de grande ampleur sur des aspects clés tels que le sexe, la race, ainsi que d'autres caractéristiques de groupes cibles.
- L'utilisation de systèmes d'IA dans des contextes spécifiques peut également conduire à l'érosion potentielle de **l'agence et l'autonomie humaines**. Les campagnes de désinformation combinées à des moteurs de recommandation élaborés, des bulles de filtrage et des chambres d'écho sont de plus en plus alimentées par une IA sophistiquée. Cela peut déclencher une dépendance et une manipulation de l'opinion.
- Le droit fondamental à la **liberté d'expression et d'information et le droit à des élections libres** peuvent être gravement affectés par les systèmes d'IA. La discrimination délibérée à l'encontre ou en faveur de certains contenus par les pratiques de filtrage et de suppression de contenus des plateformes de médias sociaux a été largement critiquée pour toucher à la fois les questions de responsabilité et de censure. En outre, le droit à des élections libres peut être mis à mal par la manipulation de diffusion d'informations sur les plateformes de médias sociaux. Des exemples marquants tels que la controverse sur les "fake news" autour de l'élection présidentielle américaine de 2016 et le scandale Cambridge Analytica/Facebook à la suite du référendum sur le Brexit montrent les impacts considérables des technologies d'IA sur la libre information et le processus démocratique.
- Les systèmes d'IA ont également un impact sur la **protection des données et le droit au respect de la vie privée et familiale**. La libre circulation des données a conduit à de nombreux conflits avec la RGPD. Dans ce contexte, une question litigieuse est liée à l'utilisation de technologies d'identification biométrique et de reconnaissance faciale alimentées par l'IA. La biométrie dite de "deuxième vague", en particulier, déploie des technologies et des algorithmes plus élaborés et collecte des données personnelles très sensibles.
- L'IA et la prise de décision automatisée utilisées par les pouvoirs publics peuvent avoir un impact sur la **bonne administration, l'accès à la justice et le droit à un procès équitable**. Il a été constaté que des domaines tels que la prédiction policière, le maintien de l'ordre public, la modélisation des risques et le système de crédit social créaient des risques pour la vie privée et la protection des données, introduisaient de nouveaux préjugés et/ou discriminaient intentionnellement les individus, par exemple en ignorant ou en discriminant ceux pour lesquels les systèmes d'aide sociale sont conçus en premier lieu. L'IA utilisée pour la prédiction policière et pour le maintien de l'ordre public est déjà opérationnelle dans plus de la moitié des États membres de l'UE, et il a été constaté qu'elle menaçait le droit de l'UE à ne pas subir d'ingérence consécutive à la collecte potentiellement illégale de données et à l'évaluation arbitraire des risques, et qu'elle soulevait des questions concernant le droit à un procès équitable et à la présomption d'innocence du défendeur.
- En outre, l'impact de l'IA sur des **groupes vulnérables spécifiques, tels que les migrants**, a fait l'objet d'un examen minutieux dans plusieurs cas.
- L'étude identifie les impacts de l'IA sur d'**autres droits fondamentaux de l'UE**, notamment la protection des consommateurs, le droit à la liberté de réunion et d'association, ainsi que la durabilité et la protection contre l'altération continue du niveau de vie des générations futures.

Cette étude examine également les **risques pour la sûreté et la sécurité** par le biais d'une analyse documentaire systématique. Nous constatons que la nature des systèmes d'IA, ainsi que le contexte dans lequel l'IA est déployée, peuvent constituer des menaces pour plusieurs aspects liés à la sûreté et la sécurité. Il est important de noter que les risques générés par l'IA peuvent apparaître à différentes phases du cycle de vie du produit, de la phase de conception et de développement de l'IA aux phases de déploiement et de post-déploiement. Nous

constatons également que les risques liés à la sécurité et à la responsabilité des produits de l'IA sont exacerbés par l'inadéquation des règles actuelles de responsabilité civile face aux systèmes d'IA. Des caractéristiques telles que la connectivité, l'opacité, la dépendance et l'autonomie en matière de données sont particulièrement importantes à cet égard. Dans des contextes spécifiques tels que les soins de santé, les considérations liées à la sécurité deviennent plus importants comptes tenus de la nature critique des systèmes à exploiter avec l'aide de l'IA. Dans de nombreux domaines, la qualité des jeux de données est essentielle à l'entrainement des systèmes d'IA : toute défaillance des données sources peut entraîner des résultats incorrects et un fonctionnement erroné tout au long de sa période d'application, invalidant ainsi l'ensemble du système d'IA. S'assurer que la source de données est digne de confiance et précise est essentiel pour prévenir les problèmes de sécurité causés par l'IA. Le biais humain dans les jeux de données d'entrainement est un problème courant dans les systèmes automatisés et peut également compromettre la sécurité d'un système d'IA.

Expérience internationale en matière de politique d'IA : cadres politiques émergents

L'étude analyse un certain nombre d'expériences internationales sélectionnées sur la conception et la mise en œuvre de cadres politiques pour l'utilisation responsable de l'IA, avec un accent particulier sur la gouvernance des risques liés à l'IA. Le paysage apparaît très hétérogène et en constante évolution : cependant, aucun pays n'a jusqu'à présent tenté d'introduire un cadre réglementaire complet et horizontal pour l'Intelligence Artificielle. Parmi les expériences observées :

Le cadre volontariste de l'**Australie** se caractérise par huit principes éthiques volontaristes en matière d'IA, en plus de conseils spécifiques aux entreprises pour la mise en œuvre. Une récente consultation publique a souligné l'importance des principes d'équité, de transparence, de contestabilité et de responsabilité, tandis que l'élaboration de normes d'IA est également une priorité nationale.

Le **Canada**, avec sa Directive sur la prise de décision automatisée, a été parmi les premiers pays à introduire un cadre réglementaire en 2018. La Directive sur la prise de décision automatisée établit une classification à six niveaux de de risque, chacun impliquant différentes exigences de niveau d'impact.

En **Allemagne**, la Commission d'éthique des données a proposé une classification des risques qui établit une échelle de "criticité" à cinq niveaux, dans une logique agnostique du secteur, permettant de classer les systèmes d'IA en fonction du degré de préjudice potentiel et proposant une interdiction totale ou partielle des systèmes classés au cinquième niveau en raison de leur potentiel de préjudice intolérable.

Les lignes directrices du **Japon** sur l'utilisation de l'IA et des données sont centrées sur les principaux défis et les questions contractuelles non résolues ainsi que sur les facteurs supplémentaires à prendre en compte dans la formulation des clauses contractuelles pour la conclusion de contrats sur les données ou les logiciels/technologies d'IA.

Le cadre de gouvernance de **Singapour** sur l'IA est un cadre volontariste et, par sa conception, neutre du point de vue des algorithmes, de la technologie, du secteur, de l'échelle et du modèle d'entreprise. Il est accompagné d'un guide de mise en œuvre et d'outils d'auto-évaluation pour les organisations. Les principaux domaines couverts par le cadre modèle sont la mise en place de structures et de mesures de gouvernance internes pour les entreprises, la détermination du niveau d'implication humaine dans la prise de décision enrichie par l'IA, la prise en compte du risque sur la probabilité et la gravité des dommages potentiels causés par

le système d'IA, ainsi que l'évaluation de l'exploitation, de la maintenance et de la gestion des données, l'interaction et la communication avec les parties prenantes, et les principes d'exclusion pour les consommateurs.

Le **Royaume-Uni** a publié un guide sur l'utilisation de l'IA dans le secteur public pour soutenir son administration afin de répondre aux besoins des utilisateurs, et pour mettre en œuvre l'IA de manière éthique, équitable et sûre. Le guide contient également des conseils sur le choix de la technique d'apprentissage automatique la plus appropriée pour un processus administratif donné. Le guide de l'Office du commissaire à l'information du Royaume-Uni sur la vérification de l'IA, en particulier, est un cadre volontariste pour les organisations couvrant les meilleures pratiques pour évaluer les risques de protection des données associés à l'utilisation des systèmes d'IA, ainsi que pour mettre en œuvre des mécanismes de responsabilité et de gouvernance ; pour garantir le respect des exigences de fond en matière de protection des données, telles que la légalité, l'équité et la transparence ; pour prendre des mesures pratiques afin de favoriser la sécurité et la minimisation des données ; et pour garantir la protection des droits individuels dans les systèmes d'IA.

Les **États-Unis** ont publié un projet d'orientation pour la Réglementation des applications de l'intelligence artificielle, censé guider l'élaboration d'approches réglementaires et non réglementaires par les agences américaines qui interagissent avec les systèmes d'IA, établissant ainsi une approche sectorielle. Le projet d'orientation vise également à réduire les obstacles à l'utilisation des technologies d'IA tout en préservant les libertés civiles. Le document comprend des détails sur la définition de l'IA, les approches d'évaluation et de gestion des risques, les moyens d'éviter une réglementation prescriptive et l'établissement de normes volontaires d'évaluation de la conformité. Au niveau de l'État, la loi sur la responsabilité algorithmique proposée par le New Jersey va un peu plus loin en obligeant certaines entités à réaliser des études d'impact sur les IA à haut risque. L'étude d'impact des systèmes de décision automatisés comprendrait plusieurs analyses coûts-avantages sur les aspects liés aux données, à la sécurité et aux risques.

Plusieurs autres cadres pertinents ont été introduits au niveau international, minilatéral ou multilatéral, ainsi que par le secteur privé. Il s'agit notamment de cadres élaborés dans la région nordique et balte, par l'UNESCO, l'OCDE, le G20, ainsi que l'Union internationale des télécommunications ou encore le Partenariat mondial sur l'IA dans le cadre du G7. Ces entités, ainsi que des acteurs non étatiques et plusieurs initiatives multipartites, façonnent le paysage mondial de la gouvernance de l'IA. L'ISO/CEI et l'IEEE Standard Association, deux importants organismes de normalisation internationaux, contribuent également à ces développements en créant des normes spécifiques à l'IA. Des instituts de recherche tels qu'AI Now contribuent à façonner le travail conceptuel sur la surveillance, la responsabilité et l'audit de l'IA, par exemple avec l'évaluation de l'impact algorithmique pour évaluer l'utilisation des systèmes d'IA dans les organismes publics.

Résumé des résultats de la consultation publique ouverte

La section 3 de l'étude présente une analyse des réponses des parties prenantes à la consultation publique ouverte sur le livre blanc de la Commission européenne sur l'IA (2020).

Dans les notes de position, le point le plus important pour de nombreux répondants était la **définition du "risque élevé"**. En effet, de nombreux répondants estiment que la définition du risque élevé dans le livre blanc n'est pas claire ou doit être améliorée (au moins 18% de tous les documents de synthèse, 74 sur 408). Beaucoup trouvent que la classification binaire en élevé vs faible est trop simpliste et certains proposent d'introduire plus de niveaux de risque.

Certains pensent que la définition est trop large, tandis que d'autres estiment qu'elle est trop restrictive.

Une autre question importante dans les notes de position était le **système de labellisation volontaire** proposé (au moins 52). Au moins 21 documents de position sont sceptiques à l'égard de la labellisation, soit parce qu'ils estiment qu'il imposera des charges réglementaires (en particulier pour les PME), soit parce qu'ils doutent de son efficacité. Dans le même temps, au moins huit documents de synthèse sont explicitement favorables à la labellisation. Les parties prenantes abordent également une série d'autres questions (voir chapitre 3).

En ce qui concerne la **définition de l'IA**, environ 15.7 % des parties prenantes indiquent qu'elles ne sont pas d'accord avec la définition de l'IA donnée par le *High Level Expert Group on AI* et la Commission. Au moins 9.3% déclarent que la définition est trop large (37), certains soulignent qu'une définition large risque d'entraîner une surrèglementation et une incertitude juridique, et n'est pas assez spécifique à l'IA. Dans le même temps, au moins 6.6 % estiment que la définition est trop restrictive (27) et peut laisser de côté certaines dimensions importantes de l'IA.

En ce qui concerne les **coûts** générés, jusqu'à 84% des documents de synthèse ne mentionnent pas explicitement les préoccupations relatives aux coûts qui pourraient être imposés par une réglementation sur l'IA (344). Dans le même temps, au moins 11% des documents de synthèse (46) mentionnent les coûts de mise en conformité en général, et au moins 7% des notes de position (29) mentionnent (également) les charges administratives.

Environ 23 % des notes de position traitent de **structures institutionnelles potentielles pour régir l'IA** dans l'UE. 10% d'entre eux sont en faveur d'une nouvelle institution au niveau de l'UE, avec au moins 6% en faveur d'une forme de nouvelle agence européenne de l'IA (24) et au moins 4% d'un comité/conseil d'administration européen moins formel (15). En même temps, au moins 3% des notes de position sont contre la création d'une nouvelle institution (14) et plusieurs mentionnent d'autres types de modalités.

Plus de la moitié des notes de position ne mentionnent pas les **exigences réglementaires proposées pour l'IA à haut risque** dans le livre blanc (surveillance humaine, données d'entrainement, tenue des données et des registres, fourniture d'informations, robustesse et exactitude). De nombreuses notes de position (au moins 23%) sont toutefois généralement d'accord avec l'approche du livre blanc concernant les exigences réglementaires pour l'IA à haut risque. Au moins 12% ne sont généralement pas d'accord, et plusieurs intervenants expriment d'autres opinions (12%).

Sur la base des réponses en texte libre aux questions ouvertes du questionnaire de consultation publique, les principales conclusions sont les suivantes :

- En ce qui concerne les **autres actions qui devraient être envisagées sur la base du livre blanc** (question ouverte - QO 1), plusieurs répondants soulignent l'importance du renforcement des compétences (58 répondants), la facilitation du partage et de l'accès aux données (21 répondants) et l'importance d'une approche globale de l'IA et de coopération ou de normes internationales (20 répondants).
- En ce qui concerne les **autres actions visant à renforcer la communauté de la recherche et de l'innovation** (QO 3), certaines réponses en texte libre font référence au réseau de centres d'excellence en recherche sur l'IA déjà existants (39) et sont sceptiques quant aux centres de recherche phares (22).
- Interrogées sur les **autres tâches des pôles d'innovation numérique spécialisés** (QO 4), certaines parties prenantes soulignent l'importance du soutien aux partenariats entre

les PME, les grandes entreprises et les universités (33) et des transferts de connaissances pour soutenir le développement de l'expertise en IA pour les PME (27).

- Interrogés sur la question de **savoir si leurs préoccupations peuvent être traitées par la législation européenne applicable** (QO 6), plusieurs répondants écrivent que la législation actuelle présente des lacunes ou qu'une nouvelle législation est nécessaire (47), tandis que d'autres pensent que davantage de recherches sont nécessaires pour exprimer une opinion (39) ou mettent en garde contre une surrèglementation (24).
- À la question de savoir si "**l'introduction de nouvelles exigences obligatoires devrait être limitée aux demandes à haut risque"** (QO 7), les répondants notent que la définition de "haut risque" n'est pas claire, que davantage de détails sont nécessaires (33) et que la distinction binaire entre haut/bas risque est trop simpliste (24).
- En ce qui concerne la question sur "**l'application/utilisation de l'IA la plus préoccupante ("à haut risque")** de votre point de vue" (QO 9), les parties prenantes ont mentionné des applications liées aux armes autonomes (41), à l'identification biométrique (34), aux applications dans les infrastructures critiques (tels que l'électricité, l'approvisionnement en eau, le nucléaire) (28) et autres.
- En ce qui concerne les **autres systèmes d'application des règles de l'IA** (QO 12), certaines parties prenantes sont favorables à des évaluations (ex ante) des droits fondamentaux (19, réponse partiellement coordonnée par des ONG), tandis que d'autres sont favorables à des auto-évaluations (14) ou à des organismes/experts externes indépendants pour assurer les évaluations (11).
- En ce qui concerne "**d'autres suggestions sur l'évaluation de la conformité**" (QO 13), certains répondants sont favorables à des évaluations externes indépendantes (32, réponse en partie coordonnée par des ONG, mais aussi par des constructeurs automobiles), et d'autres veulent éviter des procédures trop contraignantes (29).
- En ce qui concerne les "c**onsidérations supplémentaires concernant les procédures d'évaluation des risques**" (QO 15), certaines parties prenantes mentionnent que les évaluations des risques doivent être répétées en cas de changements après la mise sur le marché (16) et que des définitions plus claires de ce qui constitue, par exemple, des "changements importants" sont nécessaires (13).

Évaluation des coûts de mise en conformité générés par la proposition de réglementation sur l'IA

La section 4 contient une évaluation des coûts de mise en conformité générés par la proposition de réglementation sur l'IA. L'estimation des coûts est basée sur des méthodologies consolidées pour l'évaluation des charges administratives et des coûts substantiels de mise en conformité. Cette étude utilise en particulier le modèle de coûts développé par l'Office fédéral de la statistique du gouvernement allemand, qui présente l'avantage supplémentaire de comporter des tableaux standardisés avec des estimations de temps par activité administrative et par niveau de complexité.

L'évaluation prend en compte les cinq exigences réglementaires énumérées dans le livre blanc sur l'IA, et calcule une estimation du coût pour chaque exigence en identifiant les activités nécessaires et en attribuant à chaque activité une estimation du temps nécessaire pour accomplir la tâche. Sur la base de tableaux de coûts standardisés et de niveaux de difficulté estimés, nous avons identifié les principales activités nécessaires pour répondre à chaque exigence. Pour vérifier nos estimations, la méthodologie et les résultats de l'évaluation des coûts ont été présentés lors de deux ateliers auxquels ont participé plusieurs experts et parties prenantes du secteur; nous avons ensuite mené plusieurs discussions bilatérales avec les parties prenantes afin de nous assurer de la représentativité des chiffres retenus. De

même, les organismes d'accréditation et les organisations de normalisation ont été invités à un autre atelier pour discuter des estimations de l'équipe sur les coûts des procédures d'évaluation de la conformité. Nous avons également effectué une analyse comparative approfondie avec d'autres procédures similaires (par exemple dans le domaine de la sécurité des produits ou des dispositifs médicaux) afin de vérifier la précision de nos estimations.

Le coût estimé de mise en conformité de chaque exigence pour une "unité AI" (en supposant des coûts de développement moyens de EUR 170,000 euros) est résumé ci-dessous :

- Données d'entrainement: EUR 2,763
- Tenue de documents et dossiers: EUR 4,390
- Transmission d'informations: EUR 3,627
- Supervision humaine: EUR 7,764
- Robustesse et précision: EUR 10,733

Le coût annuel estimé de la mise en conformité de la main-d'œuvre pour un seul produit d'IA est de EUR 10,977. Si l'on ajoute l'achat de données et de services externes, ainsi que l'embauche de personnel supplémentaire, ce coût s'élève à **EUR 29,277**. Nous avons ensuite effectué une évaluation des activités qui se dérouleraient également sans intervention réglementaire supplémentaire (facteur dit "business as usual", ou BAU) ainsi que des effets d'apprentissage au fil du temps, et nous avons conclu que **les estimations de coûts pour 2025 seraient réduites d'environ 36%**.

Le coût total de conformité a ensuite été extrapolé à la taille estimée du marché et projeté dans le futur. Nous constatons que le **coût total de mise en conformité pour l'industrie mondiale de l'IA est estimé entre EUR 1,6 milliard et 3,3 milliards en 2025**, en supposant que seulement 10% des unités d'IA seront soumises aux exigences réglementaires (c'est-à-dire celles identifiées comme " à haut risque ").

La deuxième partie de l'évaluation des coûts concerne les coûts consacrés à un processus de certification d'un produit IA réglementé par le biais d'une évaluation de la conformité. Tout d'abord, nous estimons les coûts de l'évaluation de la conformité d'un produit unique (une "unité IA") dans le cadre de la procédure d'examen de l'UE, en nous appuyant à la fois sur une approche ascendante et sur une analyse comparative pour parvenir à des estimations finales des coûts. Deuxièmement, nous estimons le coût de l'évaluation de la conformité d'une unité IA dans le cadre d'un système de gestion de la qualité (SGQ). Dans le cadre de cette procédure, le coût comprend également les dépenses uniques liées à la mise en place d'un SMQ (y compris l'analyse comparative et la validation par différents experts). Nous concluons que, selon des hypothèses raisonnables, **l'obtention de la certification d'une unité d'IA par le biais de l'examen de type européen peut coûter en moyenne EUR 16,800 à 23,000, soit environ 10 à 14% du coût de développement. D'autre part, la mise en place d'un nouveau SMQ peut coûter EUR 193,000 à 330,000 au départ, plus EUR 71,400 de frais de maintenance annuels**. La plupart des coûts pourraient être partagés entre différents produits d'IA dans le cas où l'entité en question aurait développé plus d'un produit IA.

L'évaluation des coûts n'est, inévitablement, qu'un exercice estimatif compte tenu de l'incertitude liée au contenu final de la réglementation, de la part des produits d'IA qui seront qualifiés de risques élevés, de la complexité de l'industrie de l'IA et des chaînes de valeur, de la difficulté de prédire si les entreprises s'appuieront principalement sur des systèmes d'IA préformés ou développeront et déploieront l'IA en interne, entre autres. Plus précisément :

- La **complexité de l'écosystème de l'IA** constitue un défi important. Les développeurs de produits peuvent acheter un système d'IA et l'intégrer dans un autre produit. Le coût de la mise en conformité variera donc en fonction de l'obtention ou non d'une certification pour

le système acheté et si le développement du nouveau produit implique des entrées et entrainements de données supplémentaire. L'écosystème complexe impliquerait potentiellement un partage complexe des responsabilités.

- **Les organismes notifiés suffisamment équipés et qualifiés sont extrêmement rares**. Ils n'effectuent généralement pas d'examens types sur des produits contenant des logiciels dans le cadre de la réglementation des dispositifs médicaux (MDR). Le test exhaustif d'un logiciel est considéré comme impossible. En outre, il n'est pas clair si la réglementation sur l'IA exigerait un véritable audit des données d'entrainement et la tenue d'un registre, ce qui pourrait impliquer des audits supplémentaires ou réguliers étant donné que de nouvelles données d'entrainement pourraient continuer à affluer.
- Sans définir clairement les exigences et les procédures d'essai, les organismes notifiés ont du mal à **estimer le coût de l'évaluation de la conformité**. Certains coûts non récurrents, tels que la formation du personnel, les frais juridiques et les machines et équipements nécessaires, peuvent décourager les petits organismes notifiés et le coût de l'évaluation de la conformité peut donc être beaucoup plus élevé au cours des premières années d'application de la réglementation.
- Nous avons fait le choix méthodologique d'**exclure des estimations de coûts les frais liés aux conseils juridiques externes et aux honoraires des consultants**. Ce choix est fondé sur l'observation que ces coûts sont largement influencés par (i) la taille d'une entreprise et la disponibilité de l'expertise interne, (ii) la préférence de chaque entreprise, et (iii) la complexité et la rigueur des exigences réglementaires dans la réglementation proposée.

STUDY TO SUPPORT AN IMPACT ASSESSMENT OF REGULATORY REQUIREMENTS FOR ARTIFICIAL INTELLIGENCE IN EUROPE

Main text

ANALYSIS OF RISKS AND HARMS CREATED BY AI APPLICATIONS

1. The rise of AI and its risks for fundamental rights and safety

Despite offering remarkable opportunities for future growth, AI is also associated with significant risks. These are related in particular to the possibility that delegating tasks to AI systems ultimately results in a violation of individual fundamental rights, including but not limited to the right to consumer protection (Article 38 Charter of Fundamental Rights of the European Union (EU Charter), the right to non-discrimination (Article 21 EU Charter), the right to freedom of expression (Article 11 EU Charter), and the right to effective remedy and to fair trial (Article 47 EU Charter). The debate on AI has heightened in recent years, with growing awareness of the pervasive impact of AI on the functioning of our democracy and society, beyond traditional individual fundamental rights[1]. At the same time, AI integrated into more articulate systems, including smart objects and enhanced connectivity, posing new challenges for safety and security. These risks may be addressed and mitigated in various ways. In business-to-business (B2B) contexts, the contractual provisions on liability allow for optimal responsibility allocation along the value chain. In business-to-consumer (B2C) and government-to-citizen (G2C) contexts, individuals lack adequate means of agency and redress, often because they are unaware of practices (nudging and manipulation are hard to detect, for example), applicable legal provisions are lacking (e.g. liability for immaterial damage) (Wybitul and Brams, 2020), or there are sufficient knowledge and swift procedures in courts.

The following sections explore the evidence and impact of AI system deployment on fundamental rights and safety. It concludes by discussing the relationship between the specific AI techniques used, the arrangements for human oversight, and the resulting risks in the EU.

a. AI and fundamental rights: an overview

The impact of AI deployment on fundamental human rights is now one of the most intensely researched subjects in the field of AI. In addition to important academic contributions, the Council of Europe (CoE) and the European Union Agency for Fundamental Rights (FRA), regulators like the UK, French and Spanish data protection authorities, as well as dedicated non-governmental organisations (NGOs) and civil society organisations such as Access Now, Amnesty International, AlgorithmWatch, and EDRi, have analysed the use of AI and its impact on fundamental human rights. As outlined in the literature, various aspects of AI deployment determine its impact on fundamental rights, ranging from data quality (FRA, 2019) to the risks

[1] Some commentators have adopted a more nuanced approach, which distinguishes between the physical and social dimensions of AI (Yeung 2020; ELI 2020).

posed by automated decision-making. AI deployment often leads private and public organisations to sacrifice fundamental rights such as equality, non-discrimination, gender equality, privacy and data protection, and the right to a fair trial in favour of efficiency and cost-effectiveness (Misuraca, 2020)[2].

Most often, AI (and more specifically machine learning) systems were found to exacerbate bias and discrimination, depriving some end-users or groups of equal opportunities. However, **there is growing evidence that the use of AI systems can have significant impacts on virtually all fundamental rights**, some of which have already been recognised by the Court of Justice of the European Union (CJEU) as producing rights and obligations between private citizens[3]. In his 2018 report to the UN General Assembly, the UN Special Rapporteur on freedom of opinion and expression observed that 'AI tools, like all technologies, must be designed, developed and deployed so as to be consistent with the obligations of States and the responsibilities of private actors under international human rights law' (UN General Assembly, 2018). The relevance of the subject is also reflected in existing international declarations, such as the Toronto Declaration prepared by Access Now and Amnesty International, which focuses on the right to equality and non-discrimination in machine learning systems, and also in the Ethics Guidelines developed by the EU High-Level Expert Group on AI (AI HLEG), which grounds the notion of trustworthy AI in the protection of fundamental rights, as enshrined in the EU Charter, and the European Convention on Human Rights (ECHR). These rights are also enshrined in relevant international human rights law. In its guidelines, the AI HLEG clarified that, in addition to legal provisions in EU primary and secondary legislation, fundamental rights can 'also be understood as reflecting special moral entitlements of all individuals arising by virtue of their humanity, regardless of their legally binding status' (AI HLEG, 2019).

However, a recent study noted that very few national AI strategies focus on human rights, perhaps due to their prevalent focus on economic competitiveness, which often leads to settings of different priorities in public policy (Bradley et al., 2020). Of the strategies that have been adopted to date, few European examples place fundamental human rights as one of the core foundations of the overall approach to AI. Those that do include Denmark, Germany, Luxembourg, the Netherlands and Finland.

[2] A key component of the use of AI is data sharing, which has advantages (data linking, tailored interventions, better allocation of public resources, monitoring of service outcomes) and disadvantages (risk of data loss, statistical or identity disclosure, secondary usage of personal data) in the context of AI used in public administration. The JRC study found that 'whereas the expectations from the use of AI in government are high, positive impact is far from straightforward and should not be taken for granted. […] while small-scale pilot studies or experiments might be successful and the promises in case of broader adoption encouraging, providing significant efforts to ensure larger scale usage of AI inside the public sector may not be enough to accomplish the ultimate goal of sustainable take-up' (p. 5).

[3] The CJEU has spelled out that some provisions in the EU Charter may produce horizontal direct effects provided that the provision is sufficiently clear and precise, unconditional, and mandatory. See C-414/16, EU:C:2018:257, para. 57; C-68/17, EU:C:2018:696; C-569/16 and C-570/16, EU:C:2018:871; C-684/16, EU:C:2018:874; C-193/17, EU:C:2019:43.

Table 1 - Human rights in national AI strategies

Human rights mentioned	States/regional organizations[31]
The right to privacy	Australia, Belgium, China, Czech Republic, Germany, India, Italy, Luxembourg, Malta, the Netherlands, Norway, Portugal, Qatar, South Korea
The right to equality / non-discrimination	Australia, Belgium, Czech Republic, Denmark, Estonia, EU, France, Germany, Italy, Malta, the Netherlands, Norway
The right to an effective remedy	Australia (responsibility and ability to hold humans responsible), Denmark, Malta, the Netherlands
The rights to freedom of thought, expression and access to information	France, the Netherlands, Russia
The right to work	France, Russia

Source: Bradley et al. (2020)

The pervasive impact of AI deployment on human rights is found in the submissions to the public consultation on the White Paper on Artificial Intelligence[4]. Many stakeholders observed that the **impact of AI on fundamental rights goes beyond the often-mentioned bias and discrimination caused by the deployment of machine learning algorithms** and urged the European Commission to take a holistic approach in addressing the challenges posed by AI to the effectiveness of current legal frameworks, notably on fundamental rights. Existing law applies to AI systems in use, despite the concern expressed by certain stakeholders that AI systems would develop in a legislative void. For example, FRA observes that 'the (potentially) wide uptake of AI-related technologies in various sectors affects virtually all fundamental rights, from freedom of expression and information (Article 11) to good administration (Article 41)', and that 'any fundamental rights-based approach to AI should take into account the potential impact on the full range of rights, and not be limited to data protection, privacy and non-discrimination'[5].

It should be borne in mind from the outset that **violations of fundamental rights do not normally stem from the deployment of AI *per se*.** Rather, i**t is the intentional programming, and thus the conscious decision to programme and use AI systems by humans (alone or through organisations) that violate fundamental rights**. Humans decide to deploy AI for various reasons, for example, the delegation of complex tasks. This practice may result in humans or organisations encroaching on fundamental rights in various ways, whether intentionally or not. Commercial, political, or other interests may contribute to the deployment of AI violating fundamental rights, for example intentionally violating user privacy and/or deploying remote biometric identification or affect recognition systems to sell attractive advertising opportunities, or nudging users towards specific political opinions to disrupt an election process. Such goals may include efficiency or the automation of decision-

[4] FRA (2018). See the contributions of Castets et al., Yeung, CCBE, Access Now, in particular.

[5] Guild et al. (2020) confirmed in their submission that 'if unregulated or regulated ineffectively, [AI] may lead to the breach of fundamental rights, including the rights to an effective legal remedy and a fair trial, as protected within the EU by Article 47 of the Charter, Article 6 ECHR and the general principles of EU law'.

making processes (e.g. when AI systems are used to filter illegal hate speech and thereby erode freedom of expression by erring on the side of 'false positives'). Public and private organisations deploy AI systems that can unintentionally perpetuate or amplify biases that already exist in society (e.g. to decide on a prisoner's parole; to recruit personnel; or to award a creditworthiness score to a given consumer). In other cases, they may deploy a learning-based system that over time starts taking decisions that violate fundamental rights, due to changes in the environment, interaction with humans, 'black hat' manipulation or (increasing) interaction with other algorithms (e.g. Microsoft Tay's swift degeneration into hate speech)[6]. Even worse, AI applications may become powerful tools in the hands of governments determined to engage in mass surveillance to enable forms of social credit scoring, or private corporations engaging in other commercially-oriented forms of surveillance, including through the use of vocal assistants or other forms of end-user interaction.

This is a limited, non-exhaustive list of risks to fundamental rights generated by AI that have already generated significant evidence and triggered a need for action. Again, it is important to reiterate that, **rather than the technology *per se*, it is its specific use and the context in which it is deployed that determines the emergence of a given risk.** For example, generative adversarial networks (GANs) are used in the retail sector to enable a new customer experience, giving consumers the option to 'try on' clothes virtually[7], they have led artists to generate paintings 'authored' by AI (Özgen and Ekenel, 2020), they are used in machine training to augment data in case of imbalanced or insufficient datasets, and are noted for allowing the generation of 'deepfakes' (composite videos and images created with real footage that portrays fictional statements and actions), which have considerable potential to produce disinformation and even threaten political stability (Jagtap, 2020). Likewise, **when properly designed and deployed, AI solutions could also have a positive impact on fundamental rights**, for example by expanding the ability of courts to offer fair and speedy proceedings to individuals, enabling freedom of expression through a more balanced and representative news offering, offering new ways of enabling the right to private life by detecting external attempts to invade user privacy, etc.

The impact of AI systems and applications on fundamental rights (and beyond) thus appears **heavily dependent on the context and the specific use**. This suggests that any evaluation of the impact of a particular AI solution on fundamental rights would need to be carried out by the deploying entity. It also suggests two crucial additional considerations related to deployers' accountability for adopting mitigating measures in case of substantial risks of fundamental rights violations and thus related to the design of the future regulatory framework for AI.

Firstly, **the oversight arrangements (e.g. the degree of human control over the machine) should not be dictated solely by reasons of cost-efficiency**. Oversight should not be an afterthought or a mitigating measure adopted after observing the specific risks generated by a given AI system, but, rather, endogenous to risk assessment, since different types of oversight arrangements have a significant impact on the ultimate likelihood that violations of fundamental rights will be generated by specific AI applications. It seems essential to consider oversight arrangements as an important factor in determining risk, with 'human in the loop' cases being potentially less risky than entirely ADM with no human involvement. It is also important to consider that even when a human is directly involved in the decision-making process, constant interaction with AI systems that suggest decisions may lead to cognitive

[6] Paul Mason (2016) "The racist hijacking of Microsoft's chatbot shows how the internet teems with hate" https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism

[7] GANs were introduced in 2014 and were immediately recognised as a potential direction for deep learning research, especially in domains such as unsupervised and semi-supervised learning, or advanced data augmentation (see Goodfellow et al., 2014).

dependence (which occurs normally when decision makers are presented with a default option), as well as cases of de-skilling, which then hinder the possibility for the human in the loop to overturn the action suggested by the AI system. Finally, well-designed and properly enforced liability rules are of the utmost importance in triggering meaningful human-machine cooperation. Cognitive bias towards default options suggested by AI systems may be further exacerbated if any potential deviation from the suggested course of action comes with enhanced responsibility.

Secondly, regardless of the type of oversight embedded in the system, the risks generated by an AI system cannot be fully addressed by simply adopting a one-off, *ex ante* risk assessment. This is because some risks and consequent harms can be equated to 'defects' (features of the design, development and deployment choices) but others may materialise as the system interacts with the external environment and continues to adopt actions untailored to the new context or learns from biased or noisy data sources. **The need for frequent retraining or upgrades emerges in the context of both rule-based and learning-based systems**: the possibility of learning from the outside environment (as in some cases of machine learning) adds dynamism and flexibility to AI systems, but creates additional risks due to the unpredictability of the outcome of algorithmic interaction with both humans and machines. Accordingly, there is no reason to *a priori* exclude specific types of technologies from the analysis of risks generated by AI systems.

The following sections consider the evidence of cases where AI threatened fundamental rights and map future potential risks.

Bias and discrimination

The issue of bias and discrimination is potentially the most prominent and well-documented impact of AI on fundamental rights (Favaretto et al., 2019)[8]. Increasing breadth of academic literature shows widespread agreement that the (mis)use of AI can create unintentional, undesirable bias, violating fundamental rights and/or leading to outcomes and impacts that are perceived to be unfair or that are outright discriminatory. This can occur during several phases of AI development, in particular due to the possibility that bias 'creeps' into the process of data collection and cleaning, algorithmic design, testing and training, evaluation, and even post-deployment, especially where systems are regularly retrained or in the rarer cases where systems 'learn' while in use such as recommendation systems.

The literature generally distinguishes between **intentional and unintentional bias**. Intentional bias and the potentially resulting in illegal discrimination can also emerge as a result of attempts to increase the accuracy and effectiveness of algorithms. For example, search engines have to treat content differently in order to produce relevant results, and attempts to make them 'neutral' frustrates the overall purpose of their operation. In social sciences, and particularly in economics, discrimination often has a positive connotation, as it avoids cross-subsidisation and tailors products and services to individual characteristics. For example, more information on individuals' willingness to pay for a specific product may lead to different prices, thus serving a wider market. Similarly, information about the likelihood that credit will be repaid leads to more accurate algorithms and more efficient setting of interest

[8] This issue seems to be increasingly felt by citizens. A recent Equinet report recently identified that only 60% of equality body respondents were aware of a public debate within their country concerning the potential of AI to discriminate (Equinet, 2020).

rates, avoiding a situation where reliable customers pay additional interest that reflects the risk generated by other customers.

The boundary between efficiency and undesirable discrimination is, however, thin and often blurred. For example, the need to observe specific individual characteristics to achieve 'optimal' discrimination often leads to undesirable intrusions into people's private sphere, violating the right to data protection and privacy. Organisations could intentionally use proxies to discriminate based on specific features of a given population. As Kroll et al. (2016, p.682) observe 'A prejudiced decisionmaker could skew the training data or pick proxies for protected classes with the intent of generating discriminatory results'. For example, an employer could attempt to avoid the legal provisions that prohibit discrimination against pregnant women by using *ad hoc* proxies that help to infer whether or not a candidate is pregnant.

A 2018 study for the Council of Europe lists the uses and sectors in which bias and discrimination have already emerged (Ziuderveen Borgesius, 2018). They include: police and crime prevention, where systems such as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) and SyRi, and predictive policing models[9] have already been subject to heated debate; selection of employees and students, where algorithms have proven to produce discriminatory results, as in a recent case involving Amazon.[10] On online advertising, Sweeney (2013) found that Google displayed ads hinting that someone has a criminal record when people searched for African-American-sounding names. Fewer high-paying job ads were shown to women (Zuiderveen Borgesius 2018; Datta et al., 2015). Other biases include price discrimination, which in most circumstances is considered welfare-enhancing by industrial economists, to the extent that it does not amount to unlawful discrimination[11]; image search and selection, including racial discrimination in associating images, as well as image recognition by AI systems[12]; and translation tools, which were found to incorporate gender bias[13].

Considering the process of AI development and deployment, AI-driven decision-making can lead to discrimination in several ways. Barocas and Selbst (2016) distinguished five ways in

[9] Predictive policing can exacerbate bias: if a certain group of people (e.g. people of colour) are stopped and arrested disproportionately for a certain crime (which might or might not be higher in reality) and data gathered from arrests are then used for forecasting/predicting future crime rates, then predictions will lead to more disproportionate stops and arrests. See Brantingham (2018).

[10] Amazon reportedly stopped using an AI system that was alleged to bias against women to screen job applicants. The system based on historical training data and decided that male candidates were preferred to females for software developer and other technical.

[11] Angwin et al. (2015) found that a company's price differentiation practice set higher prices for people with an Asian background: 'Customers in areas with a high density of Asian residents were 1.8 times as likely to be offered higher prices, regardless of income.'

[12] Kay, Matuszek and Munson (2015) found that 'image search results for occupations slightly exaggerate gender stereotypes and portray the minority gender for an occupation less professionally. There is also a slight under-representation of women.' Moreover, some image recognition software has difficulties in recognising and analysing non-white faces. Most notable racial bias incidents include Google Photos recognising African American faces as gorillas and Nikon's digital cameras prompting a message asking 'did someone blink?' to Asian users (Zhang, 2015; Buolamwini and Gebru, 2018).

[13] Caliskan et al. (2017) gave an illustrative example of gender biases of the AI behind the automated translation tool of Google. When people typed 'He is a doctor. She is a nurse' into Google, Google Translate returned: 'O bir hemşire. O bir doktor'. Those Turkish sentences are gender-neutral, as Turkish does not differentiate between 'he' and 'she'. However, Google Translate provides: 'She is a nurse. He is a doctor'.Google has in the meantime partially updated these search results. Prates et al. (2020) tested 12 gender-neutral languages,in Google Translate, and found that Google Translate 'exhibits a strong tendency towards male defaults'. Moreover, 'male defaults are not only prominent but exaggerated in fields suggested to be troubled with gender stereotypes, such as STEM (Science, Technology, Engineering and Mathematics) jobs.'

which AI decision-making can lead, unintentionally, to discrimination (see Zuiderveen Borgesius 2018, p. 10 for a good summary):[14]

- How the 'target variable' and the 'class labels' are defined. AI systems typically look for correlations in datasets used as training data and label-discovered correlations as a 'model' or 'predictive model'. Barocas and Selbst (2016) explain that 'by exposing so-called "machine learning" algorithms to examples of the cases of interest (previously identified instances of fraud, spam, default, and poor health), the algorithm "learns" which related attributes or activities can serve as potential proxies for those qualities or outcomes of interest.' Outcomes are then called 'target variables'. The problem is that while the execution of rather mundane tasks (e.g. spam filtering, or patterns of energy consumption in an industrial setting) often creates limited concerns related to the identification and interpretation of target variables. In other circumstances, it is less obvious what the target variable should be and developers have to create new classes by establishing parameters that may end up incorporating in the data discriminatory effects on specific groups or individuals. As explained by Zuiderveen Borgesius (2018), 'suppose, for instance, that poorer people rarely live in the city centre and must travel further to their work than other employees. Therefore, poorer people are late for work more often than others because of traffic jams or problems with public transport. The company could choose "rarely being late often" as a class label to assess whether an employee is "good". But if people with an immigrant background are, on average, poorer and live further from their work, that choice of a class label would put people with an immigrant background at a disadvantage, even if they outperform other employees in other aspects.' Discrimination can thus creep into an AI system because of how an organisation defines target variables and class labels.
- How training data are labelled. AI decision-making can lead to discrimination if the training data are chosen in a way that is intentionally or unintentionally discriminatory. Again, Barocas and Selbst (2016) explain that this can occur whenever an AI system is trained on biased data and whenever the AI system learns from a biased sample. In both cases, the AI system will reproduce that bias. Examples are countless and many have been brought to the attention of the general public by widely read contributions, such as O'Neil (2016). For example, as early as the 1980s, the first cases of discrimination emerged: a medical school based in the UK which used training data on the admission files from earlier years to sort medical applications only to discover that the computer programme then discriminated against women and people with an immigrant background, reflecting and amplifying the same bias of the people that selected the students during earlier years. Similarly, a widely read ProPublica study compared two stories of prisoners awaiting parole, showing how machines (specifically, the US COMPAS algorithm generating a recidivism score) end up incorporating bias from the very outset (Angwin et al., 2016). In a similar vein, the use of Big Data and predictive policing techniques in some cities around the world has led to concerns over racial biases (Ferguson, 2017). In 2016, commentators argued that 'AI is racist' when a beauty contest that was to be decided by an algorithm, using supposed 'objective' factors such as facial symmetry and wrinkles, led to the almost total exclusion of dark-skinned contestants (Levin, 2016). The use of training data incorporating sampling bias is typical and well-documented for numerous facial recognition technologies, which perform less well with black people than with white people, and least well with black women. One reason is that the software used is

[14]The following explanation relies heavily on Barocas and Selbst (2016) and the discussion by Zuiderveen Borgesius (2018).

trained predominantly with images of white people, and relatively few images of Asian, black or brown people. The way the data are labelled also seems to be biased: an online search for 'unprofessional hair' yields search results that show a majority of black women, while a search for 'professional hair' shows white people. This effect is mainly caused by the prejudices of people who describe (label) digital images that are used as training data for machine learning applications that automatically generate descriptions of images. Equally, behavioural bias can be translated into data which then leads to discriminatory AI decision-making systems. For example, automated tools used for recruitment often rely on past recruitment patterns that are reflected in data. These data frequently contain a bias against minorities, which historically find it harder to find desirable jobs, because the behaviour reflected in the data was discriminatory. Without active measures, a tool relying on such data might discriminate. The use of behaviourally biased training data is also common within the supervised learning context, for example, AI in recruitment and human resources. Several emerging applications that potentially lead to discrimination have been observed in the use of AI by public institutions, for example in law enforcement.

- How training data are collected. Crime data are very likely to be biased as police may stop more people with an immigrant background. 'If police focus attention on certain ethnic groups and certain neighbourhoods, it is likely that police records will systematically over-represent those groups and neighbourhoods' (Lum and Isaac, 2016). The 2016 ProPublica study and the use of predictive policing practices around the world also suffer from this potential source of bias. If an AI system is fed with biased training data, it will learn that people of a specific ethnic origin, of a given provenance, or living in a specific neighbourhood are more likely to commit a crime. Predictive models using biased data are likely to produce biases (Lum and Isaac, 2016). Therefore, policing based on biased crime statistics would cause a feedback loop (Zuiderveen Borgesius, 2020). Poor people may also be under-represented in a dataset, as in the case of Street Bump. Street Bump was an app used in Boston city to collect GPS feeds to report to the city council the road conditions. However, poor people had lower uptake of the app and thus poor neighbourhoods are underrepresented and received fewer repairs (Zuiderveen Borgesius, 2018).

- Feature selection and algorithmic design. Feature selection is about simplifying the world to generate a prediction automatically. As the organization makes decisions on feature selection, it might introduce bias against certain groups of people. For example, as it is less likely that some racial groups study in famous and expensive universities, an employer selects candidates based on their education background may cause discriminatory efforts. Such effects can also stem from designer bias in automated tools, sometimes inadvertently. One example is the 'racist soap dispensers' deployed by hotels several years ago. These used visual sensors to detect a hand under the dispenser to release soap. They only worked with white hands, because of how the developers had calibrated the tool. It can be assumed that their idea of skin colour did not include darker skin types (Goethe, 2019). Organisations can therefore cause discriminatory outcomes in their selection and evaluation of features that an AI system uses for prediction.

- Proxies and redundant encodings. Some data that are included in the training set may correlate with some protected characteristics that entail discrimination. As Barocas and Selbst (2016) state, 'criteria that are genuinely relevant in making rational and well-informed decisions also happen to serve as reliable proxies for class membership '. As Ntoutsi et al. (2020) suggest, removing or ignoring sensitive characteristics may not prevent discrimination and bias as they may or may not be present in the data. Other correlated variables may be taken as proxies. For example, some neighbourhoods in US cities, which are highly correlated with a certain race, have been

> subject to systematic denial of same-day purchase delivery (Ingold and Soper, 2016). On the other hand, sensitive features in data may help design a fair model (Zliobaite and Custers, 2016). On Facebook, where users can choose to not reveal their sexual orientation, nevertheless was found to expose sexual orientations through those who publicly stated their orientations in the user's friend list (Jernigan and Mistree, 2009). 'Computer scientists have been unsure how to deal with redundant encodings in datasets. Simply withholding these variables from the data mining exercise often removes criteria that hold demonstrable and justifiable relevance to the decision at hand.' Thus '[t]he only way to ensure that decisions do not systematically disadvantage members of protected classes is to reduce the overall accuracy of all determinations ' (Barocas and Selbst, 2016, p. 721).

Against this background, bias and discrimination emerge very clearly even in the embryonic stage of the development of AI applications in various sectors, especially in commercial AI applications.[15] Without concrete safeguards during data collection and sampling, training and design of algorithms, and also in the adoption of mitigating measures during the deployment stage of the AI product, these techniques can increase discrimination, which may become more and more difficult to detect over time. Inevitably, these cases emerge most often in B2C and G2C contexts and are typical of AI techniques that are more dependent on historical data, as well as learning-based systems that observe the environment and incorporate in their decisions all the biases typical of daily life.

The responsibility for the emergence of bias and discriminatory outcomes never lies with AI systems *per se*. AI developers can mitigate the risk of discrimination in various ways: by ensuring adequate human oversight according to the specific use; by engaging in bias-aware data collection (Ntoutsi et al., 2020); by carrying out detailed and careful *ex ante* risk assessments and testing or simulations aimed at checking the fairness of the outcomes also after repeated rounds of application[16]; by using a variety of existing methods (Sánchez-Monedero et al., 2020) or dedicated AI systems to more easily detect bias. As a matter of fact, AI can not only exacerbate bias but can be deployed to prevent bias and to help developers to avoid developing systems that reproduce historic bias embedded in pre-existing human practices (Kleinberg et al., 2019).

Biases and discrimination have always existed in the job market and it is unclear whether or not AI is contributing to solving these problems. Sánchez-Monedero et al. (2020) analyse how the widely used prominent automated hiring systems (HireVue, Pymetrics and Applied) in the UK deal with bias and discrimination. The paper concludes that 'it is not clear how relevant stakeholders, not least job seekers, can access and understand information about how decisions about their eligibility might have been reached'. The authors conclude that automated hiring systems are untransparent toward assessing if internal discrimination and bias are incorporated into the algorithms' decision-making process. The researchers conclude that GDPR transparency rights make these practices incompatible with EU data and privacy legislation (also see Ntoutsi et al. 2020).

[15] Boulamwini and Gebru (2018) found that the error rate of gender classification systems for darker-skinned are the higher than lighter-skinned subjects.

[16] Discrimination risk in the field of AI decisions has been an emerging field. The organization FATML (Fairness, Accountability and Transparency) has been organizing workshops and conferences to bring together researchers and practitioners. The conference has been renamed ACM FAcct in 2020.

Human agency and autonomy: dark patterns, filter bubbles and hyper-nudging practices

The need for a human-centric approach to AI has been stated by the European Commission, as well as by several other institutions and private organisations. In executing tasks, AI can be used to exert a significant impact on human agency, including triggering cognitive manipulation through **'dark patterns' and interaction with sub-conscious processes**; generating addiction on the side of the end-user; **hyper-nudging individuals** towards specific purchasing decisions (e.g. recommendation systems) or political opinions (*Cambridge Analytica*). 'Dark patterns' are used to induce consumers to engage in purchasing activity or to give up their personal data. Dols (2020) reports that 'large tech companies … have continued to employ dark patterns to skirt GDPR Articles 5, 6, 7, 9, and 25', echoing recent reports by the Norwegian NGO, Forbrukerrådet (Kaldestatt and Myrstad, 2018). The author also finds that deep learning and the increasing use of anthropomorphic AI can lead to even subtler dark patterns, giving users the illusion of control while effectively steering their choices.

Likewise, algorithmic techniques in news selection and exposure risk generating a 'state of intellectual isolation' (an 'echo chamber'), which occurs whenever an individual interacts with highly similar news sources and other like-minded users, powered by an algorithm that only feeds users based on their perception of what they will like or be interested in (Renda, 2018). According to Negroponte and Sunstein, this 'daily me' problem is the product of behavioural biases (such as the confirmation bias, whereby we tend to like what we already agree with) (Sunstein, 2001) and the use of algorithms for personalised search, which selects content from a narrow set of sources based on users' past search activity. The problem has been acknowledged by Bill Gates, the Microsoft co-founder, who commented, 'you're not mixing and sharing and understanding other points of view' has turned out 'to be more of a problem than I, or many others, would have expected' (Joyce, 2017).[17] The European Commission notes that 'new technologies can be used, notably through social media, to disseminate disinformation on a scale and with speed and precision of targeting that is unprecedented, creating personalised information spheres and becoming powerful echo chambers for disinformation campaigns' (European Commission, 2018). The term 'echo chamber' is sometimes conflated with the more controversial concept 'filter bubble', recently subjected to criticism, when four German researchers showed that Twitter and Facebook users had a *more* varied news diet than others in their study published in the Proceedings of the National Academy of Sciences (Scharkow et al., 2020). Also a report from Oxford's Reuters Institute (Fletcher, n.d.) criticised the concept of filter bubbles in the real world.

AI systems have been found to generate problems of **addiction and opinion manipulation** for end-users (Cohen 2018). Combinations of AI algorithms and design techniques directly or indirectly aimed at directing and controlling user attention have become prevalent in social media, video and other media sites and video games. AI-powered social media with powerful recommendation systems, such as TikTok, have become extremely popular by creating echo chambers around users, just as Netflix invests billions of dollars to improve its AI-powered recommendation engine, which reportedly accounts for approximately 70% of its revenue. While these are all lawful practices within reasonable limits, they can also make users extremely vulnerable and easily deceived. Bodkin et al. (2020) report that increased screen time is implicated in teenage depression and suicide (Madhav et al., 2017) and a recent survey has shown 'how prevalent feelings of regret by users are about the apps they use - and that regret is highly correlated with the time users spend' (Centre for Humane Technology, n.d.).

[17] Also see 'Blue Feed, Red Feed' by the Wall Street Journal at http://graphics.wsj.com/blue-feed-red-feed/

More generally, the impact of AI on human agency has been subject to an extensive literature on the evolution of human-machine interaction. Sunder (2020) reveals how **humans tend to interact with computers in a similar way as with humans** (Reeves and Nass, 1996), or they become immersed in digital environments (Lombard and Ditton, 1997), or even merging with technology as a 'cyborg' (Biocca, 1997). The rise of so-called **'extended reality'**, in which virtual reality meets AI techniques such as GANs, is likely to bring new challenges for human agency, leading to a significant blurring of the boundaries between fiction and reality. The emergence of a market for 'grief bots', including the recently released South Korean documentary, showing a mother celebrating the seventh birthday of her daughter who had passed away three years before but was 'reproduced' thanks to GAN techniques, are early examples of a market that is likely to disrupt our future understanding of the line between life and death (Park, 2020).

Interference with human agency, in turn, compresses the possibility for humans to act autonomously. This occurs at various levels, whenever AI systems interact with humans offering them a default option, which ends up skewing their decision-making freedom. Linked to this phenomenon - often termed 'hyper-nudging' - are the broader issues of **'de-skilling'**, which refers to humans' tendency to under-invest in specific skills and over-rely on the accuracy and perfect functioning of machines with which they interact, and the issue of **distancing businesses from liability** thanks to the intermediation of learning-based systems (e.g. chatbots), which suggest courses of actions and nudge users (both professionals and ordinary citizens) towards specific (desired) courses of action.

All of these emerging phenomena raise the issue of how to preserve human 'control' over AI systems, and reduce the individual subjugation to algorithms designed to maximise user exploitation in commercial or political terms. Possible reactions entail the adoption of **specific approaches to meaningful human oversight**, as well as the introduction of **legal rules on (vicarious) liability** that do not create a confirmatory bias in individuals assisted by AI systems.

In recognising the impossibility of guaranteeing that a human is always in charge of final decisions on the output of an AI system (which would, in many instances, frustrate the very aim of task delegation to AI), distinctions can be drawn between cases in which human oversight of AI systems is carried out by securing a **human 'in the loop' (HITL), 'on the loop' (HOTL), or 'in command' (HIC)** (AI Law, n.d.). According to the European Commission High-Level Expert Group on AI, 'HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impacts) and the ability to decide when and how to use the system in any particular situation' [18].

Finally, in assessing the potential for AI systems to threaten human agency and autonomy, it is important to recall that the phenomena described above would also emerge in the absence of AI techniques. However, uses of AI and in particular ADM involving learning-based algorithms can produce these effects at an unprecedented scale. Despite many legal rules applicable to the deployment of AI in various sectors, the specific nature and incremental risks are not yet adequately addressed in the EU. **Legal remedies** for many of the problems illustrated above are either absent (e.g. for AI systems fostering addictive behaviour), not yet adapted to the large-scale use of AI (e.g. civil liability and consumer protection rules dealing

[18] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1

with deceptive and unfair commercial practices vis-à-vis end users covered in the Unfair Commercial Practices Directive and in the Product Liability Directive), or left to self-regulatory or co-regulatory schemes that delegate enforcement to the same players that design and deploy the algorithms.

Freedom of expression and information, and the right to free elections

In the world of over-abundant information, algorithms play an increasingly vital role in selecting, filtering, moderating, ranking and offering content to end-users. Given the volume of information and data produced daily, **there is no practical alternative to the use of AI systems to ensure that users find a relevant, personalised, lawful selection of content**. In this respect, AI-enabled algorithms have proven useful in identifying and removing content labelled as infringing copyright or containing hate speech or other illegal material. Organisations that are traditionally concerned with the protection of fundamental rights have readily acknowledged the 'enabling' nature of AI, including search engines, in providing new possibilities for freedom of expression[19]. However, the opportunities offered by automation are also accompanied by increasingly recognised risks. Concerns have been expressed both about the individual right to freedom of expression and to maintaining a pluralist, accessible and inclusive public debate.

The use of data-driven AI systems that are in charge of organising, moderating, selecting and filtering content can contribute to the polarisation of the public debate, creating less space for original content that does not fit the interests of large groups of users and leaving most of the users in self-referential echo chambers[20]. The result is a somewhat paradoxical outcome whereby 'neutral' search engines (if they can be said to exist, see Renda, 2015), inevitably end up hiding 'long tail' results and minority views and opinions - only a proactive approach to moderating content and leaving space for non-dominant voices can rebalance this trend. However, such an approach requires the exercise of editorial control and opens additional possibilities for discriminating against specific types of content, weakening freedom of expression, especially when the market structure leads to the emergence of super-dominant platforms, such as search engines (Pasquale, 2016)[21].

Issues related to freedom of expression have emerged since large general and vertical search engines (e.g. Google, YouTube) and social media platforms (e.g. Facebook, Twitter) began to adopt several **filtering and content removal practices** (Urban et al., 2016), aimed at improving customer experience and loyalty, implementing codes of conduct or even enforcing the law (as in the German copyright law). In the latter case, in particular, the need to avoid liability for failure to promptly react to the posting of unlawful content led operators to deploy AI systems that err heavily on the side of 'false positives'.

Identifying hate speech has proven to be an extremely complex task for AI and can still be rather easily gamed, as shown by research on Google's Perspective API and in real-life events, for example during the terrorist attack in Christchurch in New Zealand in March 2019,

[19] Council of Europe, Recommendation of the Committee of Ministers to member States on the protection of human rights with regard to search engines, CM/Rec(2012)3, Adopted by the Committee of Ministers on 4 April 2012 at the 1139th meeting of the Ministers' Deputies, paragraph 1. Available at https://wcd.coe.int/ViewDoc.jsp?id=1929429 (last visited on 25 September 2017).

[20] See EDPS (3/2018). Opinion on online manipulation and personal data, which identifies numerous areas of deception and manipulation caused by AI. However, empirical work on the existence and impact filter bubbles and echo chambers varies across the EU. (Nguyen et al., 2014; Zuiderveen Borgesius et al., 2016).

[21] Helberger and Trilling (2017) have compared Facebook to a 'news editor [that] has editorial responsibility for its trending topics'.

which triggered important international initiatives on curbing the use of the internet by terrorists, such as the Christchurch Call[22]. Llansó et al (2020) survey the use of natural language processing and image recognition techniques in content moderation and identify instances of false positives and false negatives[23], potential bias and algorithmic discrimination, large-scale processing of user data and profiling, and presumptions of appropriateness of prior censorship decisions. This is unsurprising in light of the limitations of machine learning identified above. Determining whether a piece of content can and should be defined as hate speech entails a delicate interpretation, balancing considerations related to hate speech and issues related to freedom of expression. **Machines in this context work with correlations and not with 'meaning', which remains inaccessible for machines. They are not good at striking this balance, despite extraordinary advances** in natural language processing (Llansó et al. 2020).

Critical cases of real-time protest monitoring using AI and facial recognition technology[24] demonstrate the tendency to adopt identification recognition technology in public spaces despite its intrusive impacts on the democratic exercise of rights to free speech and movement in public life and regardless of the violation of privacy rights.

The International Mechanisms for Promoting Freedom of Expression issued the 'Joint Declaration of Freedom of Expression and the Internet' in 2011, which states that content filters by governments or administrations, which are not end-user controlled, 'are a form of prior censorship and are not justifiable as a restriction on freedom of expression.'[25] The statement highlights the severe impact of automated filtering algorithms on the freedom of expression, mainly also because the many tools built and technical advances made still work as automated systems that ultimately act[s] as a prior restraint on speech, regardless of the accuracy of the tool used' (Llansó, 2020).

Another domain in which the increasing use of AI brings both opportunities and challenges is that of public governance, particularly the democratic process and the relationship between citizens and their administrations and the **'right to free elections'** (Council of Europe, 2017). The digital economy has brought new ways for citizens to contribute to public life, in particular by offering them ways to voice their opinions on platforms and social media and to inform themselves through an abundance of information sources. The development of AI tools to moderate and curate content can facilitate such access and contribute by helping individuals to single out the content that most interests them and navigate through the zettabytes of data available on the internet without losing their orientation. At the same time, AI use is helping administrations in the delivery of public services (Joint Research Centre, 2020), through a combination of techniques aimed at enabling citizens' rights, estimating citizens' future

[22] Recently Facebook reported having removed 9.6 million pieces of content of hate speech in 2020Q1, up from 5.7 million in 2019Q4, through a machine learning (NLP) system that detected 88.8% (8.5 million posts) before users reported them. AnA example of an NLP tool is Google/Jigsaw's Perspective API, which is an open-source toolkit that allows website operators, researchers, and others to evaluate the 'toxicity' of a post or comment through its machine learning models. However, researchers found both outstanding biases and problems of misclassification that disproportionately affects different racial groups, as well as easy ways to deceive the system. The research team behind Perspective cautions: 'We do not recommend using the API as a tool for automated moderation: the models make too many errors.'

[23] False positives and false negatives are a way to ensure statistical accuracy. 'Algorithms are often set to only report back a match if they have a certain degree of confidence in their assessment. The use of these confidence thresholds can significantly lower match rates for algorithms by forcing the system to discount correct but low-confidence matches' (Crumpler, 2020).

[24] The AI Now Report (2020, p. 11) cites cases of public surveillance with facial recognition technology in Hong Kong, Delhi, Detroit and Baltimore.

[25] The joint declaration can be found here: https://www.osce.org/fom/78309

behaviour, co-creating future public programmes with citizens, and gauging community sentiment before a programme is implemented. In most cases, the rise of large tech giants has led government officials and politicians to rely on private media outlets to communicate with citizens. During the first weeks of the COVID-19 pandemic, for example, politicians extensively communicated with citizens through tweets, press conferences on Facebook lives, or other privately-run platforms. In this context, reliance on AI tools is again inevitable, given the amount of information involved. Such use triggers severe consequences in terms of the quality and soundness of the democratic process, protection of the fundamental right to a fair trial, and, more generally, the right to good administration (see the corresponding section). This section provides a (necessarily) brief outline of the main problems in this field.

The ability of adversarial AI systems to manipulate and steer public opinion through algorithmic interaction on social networks is well-known and researched, particularly after the 'fake news' controversy around the 2016 US presidential election, the UK's Brexit referendum and the subsequent Cambridge Analytica/Facebook scandal. The COVID-19 pandemic likewise triggered substantial spreading of disinformation, leading social networks to flag some reported 50 million pieces of COVID-19 related content by April 2020. The flagging was done by an algorithm, using data from about 7,500 articles scrutinised by independent fact-checking partners. The Joint Research Centre (JRC) released a machine learning algorithm called 'Misinfo Classifier', which claims to detect misinformation up to 80% by assessing the language used in news articles.[26]

As often happens with technology, the future will see a constant race between AI tools aimed at spreading misinformation (including 'synthetic text' and deepfakes, powered by new AI solutions such as GPT-3 and GANs) and attempts by public and private institutions to counter these adversarial attacks through equally sophisticated AI tools. This dynamic effectively results in a cyberwarfare or 'AI v. AI' scenario in which humans have less control. After years in which propaganda and disinformation campaigns required large-scale investment, the landscape of disinformation is now effectively 'democratised': For example, machine learning techniques that create synthetic media has been misused to fabricate high-volume submissions to federal public comment websites in the US (Weiss, 2019). While the pre-training required to produce synthetic media can be quite costly in time, financial and human resources, the application of transfer learning is now dramatically reducing the time and effort involved. The fact that most of these models are partly released on repositories such as GitHub facilitates their repurposing for malicious use. Tully and Foster (2020) show how models such as StyleGAN2, SV2TTS and GPT-2 can be fine-tuned to generate synthetic media capable of deceiving the general public in various ways[27]. In particular, image

[26] https://ec.europa.eu/jrc/en/news/jrc-release-ai-tech-coronavirus-fact-checkers

[27] StyleGAN2, like its predecessor StyleGAN, is designed as a GAN. GANs consist of two underlying networks that are pitted against each other (hence 'adversarial') - a generator, which generates new instances of data, and a discriminator, which evaluates these instances for authenticity by deciding whether each one belongs to the actual training dataset or not. If you generate images from pre-trained StyleGAN2 off-the-shelf, it outputs random, high quality, and highly diverse images that appear in a similar orientation as the images on which it was pre-trained. These images are not present in StyleGAN2's original training set, but are completely fabricated from the generative model—these people do not exist and never have. SV2TTS is a complex, three-stage model that can perform voice cloning—or text-to-speech from arbitrary text inputs to captured reference speech in real time. SV2TTS is comprised of three underlying neural networks – first, the speaker encoder is trained on thousands of speakers in order to learn an abstract representation of human speech and squeeze it into a compressed embedding of floating point values. Then the synthesiser, which is based on Google's TacoTron2, takes text as input and returns a mel spectrogram, a numerical representation of an individual's voice. Lastly, the vocoder, based on DeepMind's WaveNet, takes the mel spectrogram and converts it into an output waveform that can be heard and comprehended. GPT-2 is an open-source neural network to predict the next word in a sentence based on previous context. The model may eventually produce fully coherent sentences and paragraphs (Rahman et al., 2020).

generation and deepfakes lead to common deceptive practices such as face-swapping, in which an autoencoder is used to analyse a large volume of images of a person to create a detailed mathematical map of the features of an individual's face (encoding) and develop a process for turning these features back into the image of the individual's face (decoding). Thus, AI used for the intentional manipulation of public opinion has become commonplace and the technical developments in AI do not suggest a halt to these practices in the short term.

Lewis and McCormick (2018) reported that a former YouTube engineer developed a program to track and collate videos that YouTube recommended in the sidebar and showed how a vertical search engine can easily be turned into a 'great radicaliser' and leading YouTube to further refine its algorithms in an attempt to counter extremism. Kaiser and Rauchfleisch (2020) reported that YouTube recommended those who watch videos of the populist right-wing party *Alternative für Deutschland* to watch videos by the more extreme National Democratic Party of Germany (NPD). Social-media analyst Ray Serrato (2018) showed that viewers searching for 'Chemnitz,' the East German city where violent anti-immigrant protects happened, were led by YouTube toward more extreme videos.

Data protection and the right to respect for private and family life

AI systems are not themselves agnostic with respect to privacy and data protection. As for all other fundamental rights, AI can be used to protect privacy and personal data, for example by detecting phishing attempts on a user email account, or by helping data subjects to manage the use of their personal data. At the same time, the ability of AI (and AI/IoT) systems to collect and analyse data facilitates old and creates new forms of privacy and data protection violations. As a result, most people believe that AI will ultimately reduce privacy[28] and the rise in data collection, reuse and repurposing has led some authors to warn of the emergence of 'surveillance capitalism' (Zuboff, 2018), with effects not too dissimilar (from the standpoint of data protection) from the massive collection and use of citizen data observed in authoritarian regimes.

This section briefly surveys the risks to privacy and data protection from the use of AI, without referring extensively to the existence of a legal framework that largely covers this specific domain in the EU (although insufficiently in relation to ADM systems), including the ECHR and the GDPR. Many scholarly papers and submissions to the public consultation on the EU White Paper on AI warned that the GDPR is insufficient protection from the widespread use of AI, in particular machine learning technologies (Hacker, 2020; Wachter, Mittelstadt and Russel, 2020). The EPRS study on the impact of the General Data Protection Regulation (GDPR) on artificial intelligence highlights a novel risk stemming from uncertainties between the interplay of the GDPR and future regulatory requirements (EPRS, 2020). The effective guidance by data protection bodies and other authorities is highlighted as essential for data controllers and data subjects to better assist companies (especially SMEs) to put in place data protection-compliant AI systems.

- ***Data aggregation***. Both rule-based AI systems and machine learning algorithms can facilitate data and image collection, processing and repurposing. These practices may impede several EU fundamental rights, including the right to private and family life and the right to data protection (Article 8) As the AI HLEG finds, if ' algorithms are used in online tracking and profiling of individuals whose browsing patterns are recorded by

[28] A survey carried out by Brookings in 2018 found that only 5% of the subjects said that they expect artificial intelligence to improve personal privacy, and almost half of them believe that it would reduce personal privacy (West, 2018).

"cookies" and similar technologies, such as digital fingerprinting, aggregated with search queries (search engines/virtual assistants) [and] Behavioural data is collected and processed from smart devices, such as location and other sensor data through apps on mobile ', this seriously undermines EU privacy and data protection principles. Similarly, the unprecedented analysis of user profiles through various datasets may lead to significant privacy and data protection breaches. However, these practices are not themselves unlawful - any justification of data collection and analysis depends on whether the data processing has a legal basis and complies with all other conditions set out in the data protection law. For instance, consent is only one of several possible legal bases to process personal data under the GDPR. However, consent must be informed and freely given and can be withdrawn at any time. This is why consent is often not the most practical legal basis for data collection and analysis. In addition, the way in which consent is obtained is often insufficient, as it does not meet the legal standard of a fully informed and freely given consent, considering the low awareness of end-users (Giannopoulou, 2020). The emergence of the IoT, with an estimated one trillion connected devices by 2035 (The Economist, 2019), will only exacerbate this problem.

- ***Data repurposing***. The repurposing of personal data is particularly problematic, as data loses its original context (Council of Europe, 2018). Repurposing of data would affect one's informational self-determination. Where the purpose of personal data processing is incompatible with the initial purpose, the processing is unlawful under the GDPR. Search engines may equally endanger the rights to privacy and data protection as current practices include large-scale data aggregation and analysis of individuals. Repurposing is facilitated by the existence of dedicated intermediaries[29]. This can become problematic also, as researchers are finding increasingly powerful open-source AI models that can be trained with repurposed data, leading to applications that were scarcely imagined by the original model developers. Recently, a creative class of data scientists has focused on extracting value and insights from datasets that are often not linked (Sareen et al., 2020) but, once combined, offer insights for gaining economic and competitive advantages particularly against rival businesses (AI Prescience, 2019). The intentional, often commercially driven 'cooking of data' (D'Ignazio and Klein, 2019, p. 162) is also facilitated by unclear legal frameworks for the trade of data. As such, the majority of datasets are not protected by intellectual property rights and no consequences arise when terms of uses are violated. Spiekermann suggests 'to specify both the concept of data ownership and exploitation claims more precisely' (2019, p. 37). However, as far as personal data are concerned, the concept of ownership rights cannot be used, as individuals have inalienable rights to their own personal data.
- ***Re-identification and de-anonymisation.*** Anonymisation techniques are one potential solution to align the hunger for data in AI models with the need to protect personal data and privacy. Such techniques imply the removal of personal identifiable information (PII). A number of anonymisation models were developed for protecting privacy such as k-anonymity, l-diversity, and t-closeness. Researchers dealing with personal data have gradually become familiar with techniques such as data masking, partial data removal, data quarantining, aggregation, data 'banding' and pseudonymisation (pseudonymisation is not the same as anonymisation and pseudonymised data remain personal data, according to the GDPR). Simultaneously,

[29] The use of massive amounts of data in machine learning can also lead to important collateral risks, which are difficult for even AI developers themselves to anticipate. For example, academic research has shown that machine learning algorithms can leak significant amounts of data and personal information used for their training (Song et al., 2017; Shokra et al., 2017), leading to further availability of personally identifiable data.

however, various ways to de-anonymise data have emerged, which frustrate almost any attempt to protect PII-data. Al Azizi et al. (2020) surveyed the state-of-the-art techniques used in de-anonymisation attacks. In recent years, numerous anonymous datasets were released and re-identified, including the medical records for 10% of the Australian population, taxi passengers in New York City, bike-sharing customers in London, subway passengers in Riga, etc. Rather strikingly, Rocher et al. (2019) developed a generative model and report that '99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes'. Their results suggest that 'even heavily sampled anonymised datasets are unlikely to satisfy the modern standards for anonymisation set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model'.

- ***Inferential analytics***. In the age of AI, most PII is not directly collected at source or consciously provided by data subjects, but, rather, is 'inferred' from the observation of user behaviour, or through the use of various types of proxies, with resulting falls in accuracy. Wachter (2019) discusses the consequences of so-called 'affinity profiling', or grouping people according to their assumed interests rather than their personal traits, which has become commonplace in the online advertising industry and many other online commerce domains. Wachter and Mittelstadt (2019) report that 'Facebook may be able to infer sexual orientation—via online behaviour or based on friends — and other protected attributes (e.g. race), political opinions and sadness and anxiety – all of these inferences are used for targeted advertising. Facebook can also infer imminent suicide attempts, while third parties have used Facebook data to infer socioeconomic status and stances on abortion. Insurers are starting to use social media data to set premiums, which is troublesome because research suggests that a person's social network can be used to draw acute and intimate inferences about one's personality: As Themistocleous et al. (2020) find, social media data can be used to infer psychological states of depression, trace and predict to some extent predict outbreaks of diseases, or make assumptions on someone's health status through its speech patterns. Wachter and Mittelstadt (2019) conclude that a new data protection right, the 'right to reasonable inferences', is needed to close the accountability gap currently posed by 'high-risk inferences'.

In summary, data aggregation, repurposing or inferential analytics pose distinct challenges to the current legislative framework. Data availability and the free flow of data across borders are key to enable various AI techniques and these exchanges can only be facilitated if compliant with the GDPR. Besides, the GDPR already stipulates how data can be processed, the purposes compatible with the initial purposes that may be lawfully pursued with personal data processing, and what can be done with personal data in terms of analytics. Such tension may be resolved by ensuring that actors are correctly informed about the GDPR and by future innovations in the domain of AI and cryptography, as well as in more decentralised models (Kaissis et al. 2020). These models include federated machine learning, differential privacy solutions (i.e. retaining the global statistical distribution of a dataset while reducing individually recognisable information), cryptographic solutions such as homomorphic encryption (i.e. an encryption scheme that allows computation on encrypted data as if it was unencrypted); secure multi-party computation; and secure hardware implementations. Kaissis et al (2020) see most promising future developments in several domains, including decentralised data processing and storing, as well as federated machine learning, efficient cryptographic and privacy primitives, machine 'un-learning' when the consent for data use is withdrawn. They also note a silver lining in the widespread implementation of security and privacy, which depends on lowering the entry barriers for researchers and developers by providing open-source tools and algorithms.

Biometric identification and facial recognition technology

Biometric identification technology is already widely used for a wide range of identification processes. While video cameras (as technical artefacts) have existed in public spaces since the 1990s, machines identifying individuals based on large-scale datasets and computing power (as human practices) is rather new. The distinction between technical artefacts and human practices should be set in a larger socioeconomic and institutional context (Lievrouw and Livingstone, 2006). Specific attention needs to be directed toward legitimate practices and the EU legal and social framework when assessing biometric identification technology and facial recognition.

Biometrics are a tool used to recognise or verify the identity of a person based on their external appearance or behavioural characteristics. This is most often done by a technological artefact, such as a video camera or a voice recognition system, which captures data. These data are then checked against a large-scale database. EU Member States use biometric features for a variety of purposes, including to identify citizens in national ID cards, passports and residence permits. However, criticism from the research community addresses several limitations of biometric identification systems. AI Now finds that 'foundational beliefs about the ability of biometric data to uniquely identify an individual are not stable and are today highly contested' (AI Now, 2020, p. 19).

The 'second wave' of biometrics deploys **more elaborate technologies and algorithms**, including 'neural wave analysis, skin luminescence, remote iris scan, advanced facial recognition, gait, speech, behavioural biometrics, and so on' (Reding, 2012, in Mordini & Tzovaras, p. 2). These 'second wave' biometrics bear new and unprecedentedly stark risks for fundamental rights, most significantly the right to privacy and non-discrimination (see discussion below). Biometrics can be programmed to infer behavioural data from video cameral material and link this to identifiable information (names). The California Consumer Privacy Act (CCPA) legally defines biometric data as 'the ability to extract an identifier template that can be algorithmically processed in order to determine whether it falls within the scope of the law' (AI Now, 2020, p. 21). In the EU, these identification techniques are regulated within Recital 51 of the GDPR: 'The processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only **when processed through a specific technical means allowing the unique identification** or authentication of a natural person.' Thus, a distinct definition of biometric data is particularly relevant to ensure that 'second wave' (and further) technologies remain legally accountable.

Biometric identification data are increasingly used to track sentimental or emotional states (emotion recognition or emotional AI). To date, inferring emotions, personality traits and other characteristics by means of biometric AI systems lacks scientific evidence. Any sentiment analysis software attempting to recognise human emotions is thus unproven. This is based on the fact that external expression does not always reflect inner emotional states accurately (Feldman Barrett et al., 2019). Alleged 'capacities' are regularly exaggerated (Varghese, 2019) and solid evidence for reliability is lacking. Strikingly, perhaps, the emotional AI market is growing significantly, from an estimated worth of USD 12 billion in 2018 to an expected USD 90 billion by 2024 (AI Now, 2019).

The company HireVue (see the corresponding section) gathers data from online video interviews, claiming to detect emotional states to predict the 'match' between applicant and company. Its AI software scans facial expressions, voice and body language, claiming to determine how suitable the person is for the role. HireVue also promises to estimate the success of a candidate in the new role. The system is said to lower recruiting costs by speeding up the hiring process. According to the Washington Post, over 100 companies have

already used HireVue to assess more than one million applicants, including in the EU[30]. However, such companies often fail to demonstrate scientifically verifiable results (Thiel, 2019).

More specifically, the use of sentiment analysis software in the hiring process entails several critical issues in the European context. Limited training data, biased decision recommendations due to the processing of previous applicant data, as well as significant disadvantages for non-native speakers and disabled people are among the most frequently raised criticisms of HireVue (Engler, 2019; MIT Tech Review, 2019)[31]. This is why numerous research institutions and civil society organisations argue that the use of emotion detection applications should be banned in hiring decisions (Vaas, 2019). Reacting to the public consultation on the AI White Paper of the European Commission (see Chapter 2), civil society organisations such as AccessNow and EDRi stated that emotion recognition systems and other tools based on doubtful science should be prohibited.

Facial recognition systems

The key technology to enable sentiment analysis and one of the most frequently used biometric techniques are facial recognition systems. Facial recognition technologies and the linked biometric data have considerable potential but also pose considerable risk to fundamental rights, particularly the right to privacy and to non-discrimination. The US cities of San Francisco (Barber, 2019b), Boston (Owaida, 2020) and most recently Portland (Hunton Andrews Kurth, 2020) have banned facial recognition technology in public spaces.

Portland City Council justified the facial recognition ban based on 'documented instances of gender and racial bias in facial recognition technology, and the fact that marginalized communities have been subject to 'over surveillance and [the] disparate and detrimental impact of the use of surveillance' (Hunton Andrews Kurth, 2020). The outright ban on any facial recognition technology in US cities points to the rapid development in this technological field and its broad impact on citizens.

While the ban was welcomed by civil society organisations, other stakeholders criticised the regulators for having 'seized the opportunity to act in the AI space—proposing and passing outright bans on the use of facial recognition technology with no margin for discretion or use case testing…' (Gibson Dunn, 2020, p.18)[32]. To date, however, the majority of development and testing of facial recognition systems is undertaken by private companies and the scientific research community is less involved.

The AI Now 'Regulating Biometrics' report (2020) details the shortcomings of facial recognition technology, making it ill-suited to replace fingerprints, for example, for identification processes. It notes that face recognition still performs poorly in applied contexts, including high error rates for '[b]lack women, gender minorities, young and old people, members of the disabled community, and manual labourers' (p. 9). As any facial recognition system relies heavily on labeled data, it is problematic that 'much of this data labeling work, often contingent and

[30] For example, the European Investment Bank has used HireVue in its hirinng application process, https://edps.europa.eu/sites/edp/files/publication/15-11-24_notification_for_recruitment_processing_operations_eib_en.pdf

[31] Rights group files federal complaint against AI-hiring firm HireVue, citing 'unfair and deceptive' practices. (n.d.). Washington Post. Available at: https://www.washingtonpost.com/technology/2019/11/06/prominent-rights-group-files-federal-complaint-against-ai-hiring-firm-hirevue-citing-unfair-deceptive-practices/

[32] https://www.gibsondunn.com/wp-content/uploads/2020/02/2019-artificial-intelligence-and-automated-systems-annual-legal-review.pdf, p. 18.

underpaid, is outsourced to firms across the world' (p. 8). The unreliably labelled data are then processed by algorithms, both supervised and unsupervised machine learning systems, to predict the match with an image within a database. The underlying assumption 'is that a strong connection exists between bodily traits and identity, and that biometric identifiers can be uniquely attributed to a particular individual…These claims of accuracy and efficiency are often taken as a given' (ibid.). This relates to the problem of statistical accuracy and false positives or false negatives depending on the facial recognition use case[33]. Without a human reviewing the results, 'higher miss rates may be preferable to allowing false positives, and strict confidence thresholds should be applied to prevent adverse impacts. However, when facial recognition is used for what is often termed investigation—simply returning a list of possible candidates for human operators to review—confidence thresholds are usually reduced, as humans are checking the results and making the final decision about how to use the information that is returned' (Crumpler, 2020).

AI Now also critiques the lack of public accountability in governmental use of facial recognition. While the public sector represents the largest customer group, the development, marketing and maintenance of its systems are outsourced to private firms (BusinessWire, 2019). This is crucial for cases 'in which facial recognition has resulted in misidentification of suspects, including cases where facial recognition is used as primary evidence to determine guilt' (AI Now, p. 11). Beyond the governmental deployment of facial recognition, face scanning practices during a music concert – without the explicit consent of attendees (Stanley, 2018)[34] – raises questions about the accountability, transparency and justification of facial recognition.

More generally, the analysis of facial recognition data is profoundly connected to personal feelings, intimate behaviours and private thoughts. The barriers to sharing these intimate data are likely to be significantly higher than sharing gender or age (see the corresponding section). It often remains unclear to users when data are collected and what type of data is aggregated and processed. This results in **power imbalances between disproportionately powerful people or companies over individual users or marginalised groups**. It also results in a lack of accountability and little or no means of challenging data collection and the resulting decisions. Users often do not have any means of redress. Less than two in 10 Europeans want to share their biometric data with public authorities (FRA, 2020)[35]. According to a comprehensive study on 'soft biometrics', half of all UK citizens did not agree to their emotional data being collected, especially given the lack of effective objection to data collection (McStay, 2020). **Facial recognition for emotional AI, in particular, obscures when and what data are collected and prevents users from exerting meaningful active human agency (Fanni et al., 2020).**

A comparative study assessing the reliability of facial recognition systems for emotion detection finds that humans are still better at recognising emotions than automatic classification (Dupré et al., 2020). The accuracy between the performance of eight emotion

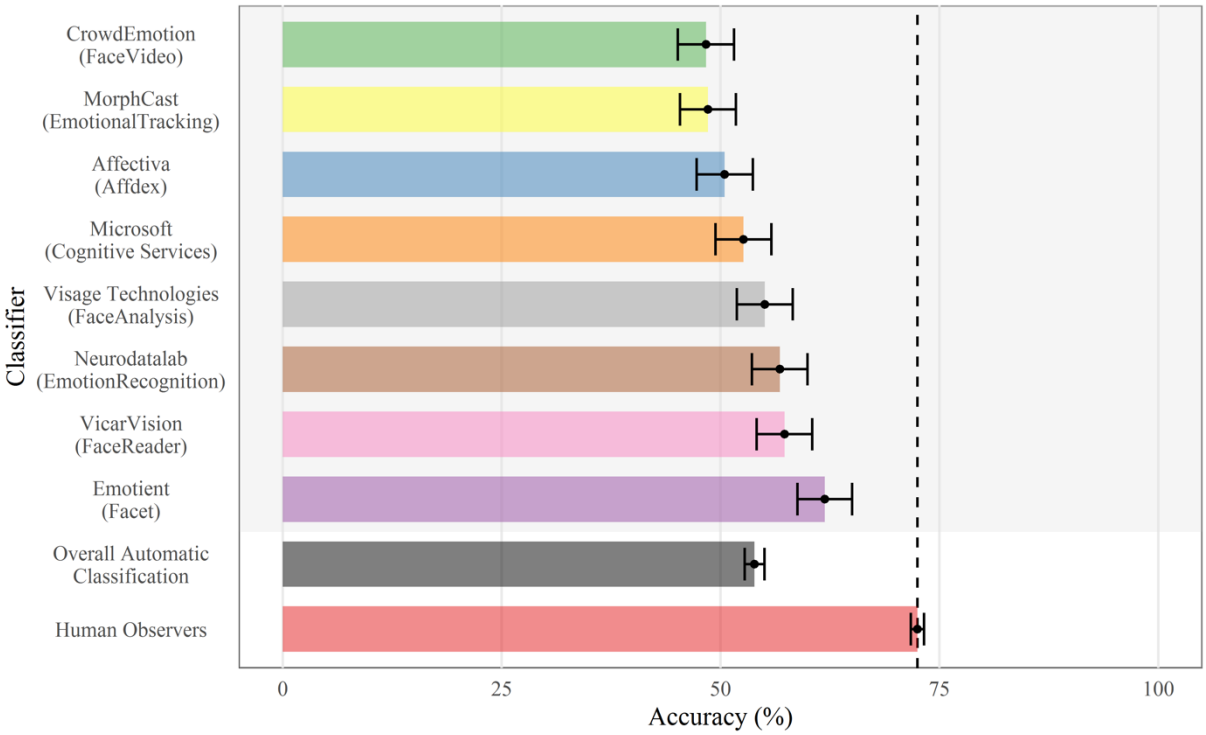
[33] According to Crumpler (2020), considering the effect on accuracy when deploying algorithms is central to avoiding so-called 'false positives': 'With facial recognition likely to be used in contexts where the user will want to minimise the risk of mistakenly identifying the wrong person (e.g. where law enforcement uses the technology to identify suspects), algorithms are often set to only report a match with a certain degree of confidence. The use of these confidence thresholds can significantly lower match rates for algorithms by forcing the system to discount correct but low-confidence matches. For example, one indicative set of algorithms tested under the Facial Recognition Vendor Test (FRVT) had an average miss rate of 4.7% on photos 'from the wild' when matching without any confidence threshold. Once a threshold was imposed requiring the algorithm to only return a result if it was 99% certain of its finding, the miss rate jumped to 35%. This means that in around 30% of cases, the algorithm identified the correct individual but did so at below 99% confidence, and so reported no match.'

[34] See https://www.wsj.com/articles/facial-recognitions-next-big-play-the-sports- stadium-11596290400

[35] https://fra.europa.eu/en/project/2015/fundamental-rights-survey

recognition AI classifiers varied between 48% and 62% while humans identified around 75% of the classified emotions (Figure 2).

Figure 1 - Mean True Positive recognition performance of automatic classifiers.



Source: Dupré et al., 2020.

The US National Institute for Standards and Technology (NIST) Facial Recognition Vendor Test (FRVT) found that the error rate for a widely deployed algorithm increased from 0.1% to 9.3% for pictures that were taken in a real-life setting. Ageing can also increase the error rate of facial recognition technology (Crumpler, 2020). Feldman Barrett et al. (2019) draw attention to a more fundamental question: face movements do not necessarily correlate with the expression of various kinds of emotions and sentimental information for everybody. Likewise, the perception of emotions and expressions, especially across cultures and social groups, is insufficiently explored and the scientific grounds for any interference between facial or sentiment analysis and assumed behaviour is rather weak. **Emotion recognition AI contains significant shortcomings in classifying and measuring emotions, as the services lack scientific reliability and validity** in decoding and interpreting emotional states or behaviour.

The insufficient scientific validation of emotion detection AI has led to calls for the prohibition of emotion recognition AI deployment. Clifford (2019) states that 'such commercial purposes [of emotion detection AI] could be banned *ex ante* considering overlaps between the EU data protection, privacy, and consumer protection frameworks.' Annex 1 provides a non-exhaustive overview of other instances where fundamental rights are infringed by AI systems.

Remote biometric identification

Another contested use of facial recognition AI systems is the remote biometric identification (RBI) of individuals. RBI systems claim to identify an individual based on physical, physiological or behavioural characteristics. These systems often operate in the background, thus providing insufficient information to individuals, who are not asked to consent to their data

being collected and processed. RBI has been widely criticised by numerous stakeholders, including digital rights organisations, civil society, politicians, and scientists (see EDRi (n.d.) for a list of articles and documents on the issue of facial and biometric recognition). Several reports state that RBI threatens fundamental rights. Human dignity, including the right to self-determination, may not be fully exercised if RBI systems were to autonomously capture data in public spaces. The normalisation of RBI used for surveillance in public spaces exacerbates discrimination and bias (see the corresponding section) and explicit consent is almost impossible to gather (EDRi, 2020). The erosion of privacy is especially concerning given that individuals often cannot object to their faces being scanned (see the corresponding section). In addition, facial recognition used for RBI 'is not only an issue of privacy, but it's also an issue of democracy in itself and pertains to self-determination. All the social problems that this software ought to solve -transnational corporate crime, violent acts ─ require social intervention. … The safety benefit is hypothetical, the feeling of surveillance is tangible in the discourse….' (Eireiner, 2020, p. 13). AI used for RBI in public spaces likely violates the essence of the right to privacy**.** RBI also raises serious questions about the GDPR necessity and proportionality principles for collecting data. In this context, the settled case-law of the CJEU confirms that '[a]n objective of general interest—such as crime prevention or public security—is not, in itself, sufficient to justify an interference [with a Charter right]' (CCDCOE, n.d.)[36]. This means that hypothetical claims to increase efficiency, enforce law or protect national security by deploying RBI are insufficient to justify the violation of EU fundamental rights. To conclude, the costs to both individual fundamental rights and democratic values far outweigh the perceived benefits of deploying RBI. As EU data protection legislation already severely restrains the processing of biometric data for identification purposes remote biometric-identification can only be allowed in very few circumstances with substantial public interest and complying with EU and national law, and with a justified, proportionate and safe use.

Several organisations have long questioned the ongoing use of RBI. The European Data Protection Supervisor (EDPS) issued a very critical statement arguing for a moratorium of deploying RBI as well as biometric data (EDPS, 2020). According to the EDPS, the adoption of AI is insufficiently scrutinised considering the wide range of impacts on individuals and on our society: 'We support the idea of a moratorium on automated recognition in public spaces of human features in the EU, of faces but also and importantly of gait, fingerprints, DNA, voice, keystrokes and other biometric or behavioural signals.' In addition, a European Citizen Initiative (ECI), 'Civil society initiative for a ban on biometric mass surveillance practices' has called for the Commission to propose legislation 'to permanently end indiscriminate and arbitrarily-targeted uses of biometric data in ways which can lead to mass surveillance or any undue interference with fundamental rights'[37].

The artefacts enabling biometric data collection do not violate EU fundamental rights, per se. Rather, it is the practices, such as algorithmic design, choices and contexts in which biometric data, facial recognition and RBI systems are deployed, that are questionable.

b. AI and ADM in government: good administration, access to justice and fair trial

Many public administrations around the world are turning towards AI solutions to improve the effectiveness and efficiency of public services, tailor their information exchange with citizens,

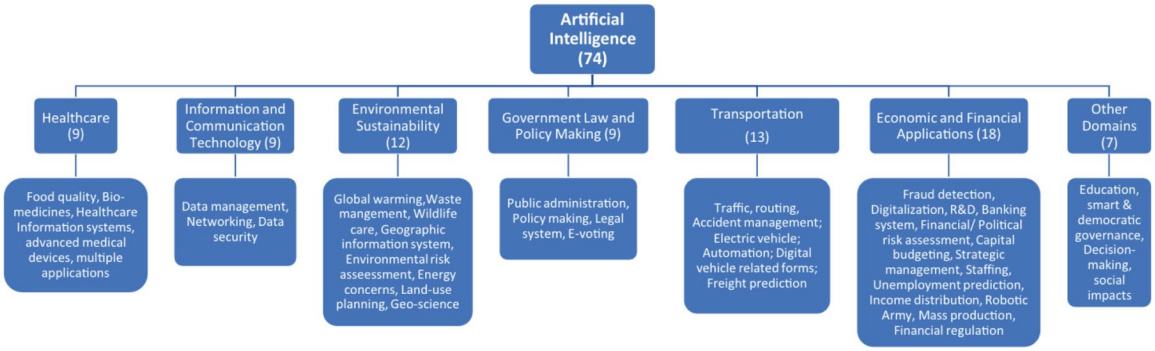
[36] https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-facial-recognition-technology-focus-paper-1_en.pdf, p. 21.

[37] If the ECI receives one million statements of support in one year (from at least seven EU Member States), the Commission will have to react by either following the request and proposing legislation or not (https://ec.europa.eu/commission/presscorner/detail/en/ip_21_22).

engage in predictive analytics, and support decision-making. This is a growing field of academic research and emerging cases, which portrays the typical dual nature of AI, as both opportunity and challenge. Given the nature of this survey, it will focus primarily on the risks generated by these emerging practices, but it should be borne in mind that the use of trustworthy AI/ADM solutions in government offers public administrations important new possibilities. Nevertheless, esteemed institutions have voiced significant concerns about the uncontrolled diffusion of algorithms in public administrations, with the Council of Europe observing that algorithmic decision-making 'is threatening to disrupt the very concept of human rights as protective shields against state interference' (Schulz et al., 2017, p.33).

De Souza et al. (2019) observe that only roughly 4% (59) of the articles published between 2000 and 2019 discussed applications of AI in the public sector. Sharma et al. (2020) find a slightly higher number of papers (74) in Web of Science and offer an organising framework for the most common uses of AI in government (see Figure 3).

Figure 2 - Mapping uses of AI in government



Source: Sharma et al. (2020)

Misuraca and van Noordt (2020) address the gap in the literature by focusing on the current state of AI deployment in the public sector, collecting 230 initiatives using AI in public services (broadly defined, see below) across the EU and observing that most of the academic research to date has focused on private sector uses. Table 3 presents a snapshot of their findings, showing types of AII techniques used, description of task executed and examples. Both rule-based systems and learning-based systems are deployed by public administrations for a variety of uses, with important findings in respect of the relative diffusion of chatbots and virtual assistants (52 cases), predictive analytics (37 cases) computer vision and ADM (29 cases). They also find that the **COVID-19 pandemic stimulated and accelerated the development and adoption of AI technologies**, which include medical applications (Bullock et al., 2020; Wang and Tang, 2020) and social distancing enforcement (Naudé, 2020).

The AlgorithmWatch report (2020) cites several critical cases of AI deployment in the EU. In Poland, the *Kwarantanna domowa*[38] app must be downloaded and uses geolocation and face recognition technology to monitor if infected people stay at home. A similar system is deployed in Hungary[39], but on a voluntary basis. In Norway, the contact tracing app *Smittestopp* was suspended after the Data Protection Authority issued a warning because it disproportionately infringed users' privacy. The Lithuanian tracing app was similarly suspended for failing to comply with the GDPR (Pugh, 2020. In Liechtenstein, people were given a 'biometric bracelet

[38] https://www.gov.pl/web/koronawirus/kwarantanna-domowa

[39] https://index.hu/belfold/2020/05/05/koronavirus_magyarorszagon_hazi_karanten_nyomkoveto_magyar_kozlony/

to collect "vital bodily metrics including skin temperature, breathing rate and heart rate"', despite numerous concerns about the effectiveness of wearables in containing the virus. More generally, contact tracing apps were not subject to *ex post* scrutiny or key performance indicators (KPIs) (AlgorithmWatch, 2020). Experts concluded that most fever cameras have 'an accuracy of +/- 2 degrees Celsius [so] the problem of false positives cannot be ignored. False positives carry the very real risk of involuntary quarantines and/or harassment', which means that citizens are unduly discriminated against because of technical inaccuracies (Electronic Frontier Foundation, in AlgorithmWatch, 2020, p. 14). Lehuedé, Filimonov and Higgins (2020) highlight that the digital infrastructures, apps and devices 'that have become a fundamental piece of the response to the COVID-19 are not subject to public accountability because they respond to the interests of economically and politically powerful transnational companies […] taking advantage of the current situation in order to gain control of services that were previously provided by the state.' Thus, several cases of AI and machine learning used in the healthcare context call for more research and oversight, especially if deployed by public authorities.

Alongside the advantages, a number of concerns about possible downsides and misuses of AI are reported, including 'black box' problems (i.e. lack of transparency and/or predictability in the inner working of the algorithms used)[40], the amplification of biases of which users might be unaware (Wirtz et al., 2019), and the **weakening of privacy protection** 'due to the fact that many devices and services gather data without the user's full understanding of what is done with it afterwards' (Wirtz et al., 2019).

[40] The FAT/ML Research community established 'Principles for Accountable Algorithms and a Social Impact Statement for Algorithms' (https://www.fatml.org/resources/principles-for-accountable-algorithms).

Table 2 - AI in government: current and prospective technologies and uses

AI typology	Description	Example	No. of cases
Audio Processing	These AI applications are capable of detecting and recognizing sound, music and other audio inputs, including speech, thus enabling the recognition of voices and transcription of spoken words.	Corti in Denmark is used to process the audio of emergency calls in order to detect whether the caller could have a cardiac arrest	8
Chatbots, Intelligent Digital Assistants, Virtual Agents and Recommendation Systems	This AI typology includes virtualised assistants or online 'bots' currently used in not only to provide generic advice but also behaviour related recommendations to users.	In Latvia, the Chatbot UNA is used to help answer frequently asked questions regarding the process of registering a company	52
Cognitive Robotics, Process Automation and Connected and Automated Vehicles	The common trait of these AI technologies is process automation, which can be achieved through robotized hardware or software	The use of self-driving snowploughs in an airport in Norway in order to improve the clearing of snow on runways.	16
Computer Vision and Identity Recognition	AI applications from this list category use some form of image, video or facial recognition to gain information on the external environment and/or the identity of specific persons or objects.	In Estonia, the SATIKAS system is in used which is capable of detecting mowed (or the lack of mowed) grasslands on satellite imagery	29
Expert and Rule-based Systems, Algorithmic Decision Making	The reason why these apparently distant AI developments are joined into a single application is their prevalent orientation to facilitate or fully automate decision making processes of potential relevance not only to the private but also to the public sector.	Nursery child recruitment system used in Warsaw. The algorithm considers data provided by parents during the registration, calculates the score and automatically assigns children into individual nurseries.	29
AI-empowered Knowledge Management	The common element here is the underlying capacity of embedded AI to create a searchable collection of case descriptions, texts and other insights to be shared with experts for further analysis.	In Slovakia, an AI system is used in the government to assist in the browsing and finding of relevant semantic data	12
Machine Learning, Deep Learning	While almost all the other categories of AI use some form of Machine Learning, this residual category refers to AI solutions which are not suitable for the other classifications.	In Czechia, AI is used in social services to facilitate citizens to stay in their natural environment for as long as possible	17
Natural Language Processing, Text Mining and Speech Analytics	These AI applications are capable of recognising and analysing speech, written text and communicate back.	In Dublin, an AI system analyses citizen opinions in the Dublin Region for an overview of their most pressing concerns by analysing local twitter tweets with various algorithms.	19
Predictive Analytics, Simulation and Data Visualisation	These AI solutions learn from large datasets to identify patterns in the data that are consequently used to visualise, simulate or predict new configurations.	Since 2012, the Zurich City Police have been using software that predicts burglaries. Based on these predictions, police could be forwarded to check these areas and limit burglaries from happening.	37
Security Analytics and Threat Intelligence	These refer to AI systems which are tasked with analysing and monitoring security information and to prevent or detect malicious activities.	In the Norwegian National Security Authority a new system is used based on machine learning is enabling the automatic analysis of any malware detected to improve cybersecurity	11

Source: Misuraca and Noordt (2020)

While many of the cases do not raise fundamental rights concerns, a significant number of ongoing initiatives could lead to the compression of citizens' privacy, the right to a private life, and the right not to be discriminated against.

Risk prediction, risk modelling and social scoring

One of the most recurring important domains of AI application is **risk prediction**. The Equinet report (Allen and Masters, 2020) identified a number of use cases for AI deployment, including assessing the risk of a person remaining unemployed, requiring care, that a child might need welfare services, crime, hospitalisation, committing fraud, and re-offending. Among the models surveyed used is 'Risk-Based Verification' (RBV), used *ex lege* in the UK by local

authorities to determine an individual's eligibility for housing and council tax benefits. Masters and Allen (2020) explain that the RBV model assigns a risk rating to applicants, based on which the level of identity verification is defined. This way, the authorities are able to better allocate resources.

Similar patterns of predicting risks through AI emerged in the context of the well-known *Systeem Risico Indicatie* **or SyRI model** used in the Netherlands to determine the risk of fraud for social security/welfare. SyRI was able to link a large number of government datasets and analyse them anonymously by the Dutch administration for generating so-called risk reports. However, the government did not provide any information as to which datasets had been combined or on the functioning of the algorithm. No (algorithmic) impact assessment was carried out before the system was used for a specific purpose. The system was eventually used to target and analyse the data of residents in low-income areas, such as certain districts of Rotterdam. Several claimants started legal proceedings against this use of the system and its underpinning law. In February 2020, the District Court of the Hague ruled 'that the Netherlands as a party to the ECHR has a special responsibility when applying new technologies to strike the right balance between the benefits the use of such technologies brings as regards preventing and combating fraud on the one hand, and the potential interference with the exercise of the right to respect for private life through such use on the other hand. From the viewpoint of protection of the right to respect for private life, which includes the protection of personal data, legislation must offer a sufficiently effective framework which allows the weighing of all interests in question in a transparent and verifiable manner.'[41] The Court thus confirmed that every state authority has a special responsibility to safeguard the right to respect of private life when it is regulating new technologies. The Court ruled that SyRI served a legitimate purpose (preventing the misuse of public funds), but the system and underlying legislation lacked fair balance and SyRI violated the right to private life. More specifically, the Court was of the opinion 'that the SyRI legislation contains insufficient safeguards to protect the right to respect for private life in relation to the risk indicators and the risk model which can be used in [the] concrete SyRI project.'[42] This confirmed that transparency is a key requirement if an application is not to fall foul of Article 8 ECHR. The Court also confirmed that SyRI created potential discriminatory effects, as it applied to so-called 'problem districts'.

Another important example is that of **profiling or credit scoring systems**. In the Danish city of Gladsaxe, for instance, a tracing tool was introduced as part of the country's 'ghetto plan' in January 2018 to detect children in vulnerable circumstances at an early stage. Municipalities were allowed to collect and combine information on children from different public sources and to categorise it according to specific risk indicators. The system then assigned a score to the family based on information such as attendance at doctor's appointments, employment and family status, mental health, and similar criteria. In December 2018, the Gladsaxe municipality was subject to a leak, which exposed the personal data of more than 20,000 citizens, including gender, age, welfare benefits and the family's special conditions. This case exemplifies the typical implications that come with profiling - not only do such programmes expose significant privacy and data protection risks, they also may be used in a discriminatory way. Most people were not even aware that they had been subject to the programme and were thus prevented

[41] ECHR cited in Nederlands Juristen Comité voor de Mensenrechten (NJCM) Consultation EU White paper on Artificial Intelligence, https://njcm.nl/wp-content/uploads/2020/06/Outline-reactie-internetconsultatie-AI-2.pdf

[42] ECHR cited in Nederlands Juristen Comité voor de Mensenrechten (NJCM) Consultation EU White paper on Artificial Intelligence, https://njcm.nl/wp-content/uploads/2020/06/Outline-reactie-internetconsultatie-AI-2.pdf

from objecting to the programme or their inclusion in it. As stated by a report on consumer credit data in the retail financial markets in the EU, credit scoring 'has been subject to several criticisms for its numerous fallacies, particularly for introducing new biases, or for making assumptions that lack universal acceptance or that may work on large numbers but not for individual cases' (Ferretti, 2017). The authors warn of personal data becoming crucial to 'the economic and social life of people determining, inter alia, access conditions to services.' These assessments highlight the further need to assess whether consumer data should be used for these important decisions, and how to ensure that personal data processing is in line with the EU data protection legislation. AlgorithmWatch (2019) reports several other cases of **personal scoring** in the EU. These include projects undertaken in Trelleborg, Sweden, where an algorithm gathers data from several databases (e.g. tax agency, bureau for housing support) and decides whether or not applicants can receive social benefits, in France, where intelligence services deployed algorithms that detect anomalous behaviour from internet users, and in Spain, where an algorithm decides if tenants are eligible to subsidised electricity prices using income and rent data (Belmonte, 2019).

A related domain in which the use of AI to support public authorities raises important concerns is **predictive policing**. First applied in the state of California based on early software developed by Jeff Brantingham at UCLA ('PredPol'), it is now a reality in many EU Member States. However, these systems are often based on proxies and algorithm variables that include criminal history and family background, which can make the past behaviour of a criminal group determine the fate of an individual. In the domain of criminal justice, different individuals will be subject to completely different treatment by public authorities. This is even worse in those countries where historical data incorporate generations of discrimination and racial bias, manifest for example in disproportionate policing of vulnerable populations (Richardson et al., 2019).

Researchers have shown that systems appearing not to use any personal data can have harmful impacts as they use proxies that can lead to similarly discriminatory – and possibly less accurate and thus more unfair – results. Ethnic or social profiling of social groups based on their neighbourhood have already appeared in the context of area-based risk prognoses, resulting in unjustified increased policing in those areas (Datenethikkommission, 2019). The Council of Europe Commissioner for Human Rights warned that 'Member States should apply the highest level of scrutiny when using AI systems in the context of law enforcement, especially when engaging in methods such as predictive or preventive policing' (Council of Europe Commissioner for Human Rights, 2019b). Gstrein et al. (2019) provide an international empirical investigation of predictive policing, reviewing established systems such as the Dutch Crime Anticipation System (CAS) released in 2013 and the PreMap project deployed in a German state to predict domestic burglary based on historic crime data ranging from 2008 to 2013. The researchers conclude that individual and group privacy rights could be significantly violated, and cast doubt on the effectiveness of predictive policing intended to reduce crime rates. Williams (2018) describes other tools such as the 'Gangs Matrix' used in London, which reportedly displays significant racial bias, and other similar tools used in Spain, France, Portugal, Denmark and Sweden to police youth gangs. Learning algorithms have already been used in predictive policing, where they help to evaluate the risk of crime through predictions of future behaviour (Kouziokas, 2017). For example, the Hesse police force has partnered with Palantir (a controversial company developing surveillance and intelligence software based on AI) to carry out some of its investigations (Monroy, 2019. Among Palantir's databases is the file 'personalised evidence' (*Personengebundene Hinweise*), which uses disputed labels such as 'behavioural disorder', 'risk of infection' or 'willingness to use violence'. The Zurich police rely on the Dynamic Risk Assessment Systems software (DyRiAS) in predictive policing. AI in policing was used in a pilot project involving facial recognition software in a Berlin train station (Finck, 2020). A recent report by AlgorithmWatch cites 14 cases of automatic image analysis from surveillance cameras using computer vision techniques in Belgium, Czechia, Germany,

Spain, France and Poland (Kayser-Bril, 2020). In Germany, the federal states Bayern, Bavaria, Baden-Württemberg, Hesse, Berlin, Northrhine-Westphalia and Lower Saxony deploy different predictive policing software to predict repetitive burglary (Heitmüller, 2019).

Figure 3 - Overview of predictive policing deployment in the EU

country	city	vendor	source	comment
	Brussels	One Télécom	source	Detection of illegal trash dumps, theft.
	Kortrijk	BriefCam	source	
	Prostějov	BriefCam	source	
	Prague	BriefCam	source	Tender in process.
	Mannheim	Frauenhofer IOSB	source	Detection of body movements that constitute assault.
	Marbella	Avigilon	source	
	Nîmes	BriefCam	source	
	Nice	Two-I	source	Not implemented yet.
	Cannes	Datakalab	source	Detects if pedestrians wear face masks.
	Roubaix	BriefCam	source	
	Marseille	Snef	source	
	Toulouse	IBM	source	
	Yvelines		source	Surveillance of high schools and one fire station.
	Warsaw	BriefCam	source	

Source: Kayser-Bril (2020)

AI used for predictive analytics systems and techniques is applied in **law enforcement, beyond policing**. Examples include the Harm Assessment Reduction Tool (HART), developed by Durham Constabulary and the University of Cambridge in 2015-16 that aims to identify people who have a moderate risk of recidivism and thus they could be given 'out of court disposal'. The idea is to reduce the number of individuals entering the justice system, and hopefully to reduce the number re-entering it (The Law Society of England and Wales, 2019). The increased use of AI systems within criminal justice and law enforcement threatens the EU right to be free from interference. In this respect, any indicative data informing such risk-scoring systems may have been collected unlawfully and thus risk scoring can happen on an arbitrary basis. Further, the right to a fair trial and innocence of the defendant interferes with low-/high risk ratings. This is particularly critical in instances where individuals are denied bail or are proven guilty despite not knowing the reasons for such sentences, as these have been determined by a 'black box algorithm' (Access Now, 2018).

Use of AI in courts and law firms: the right to a fair trial

In several countries, courts experience a lack of resources and a significant backlog, which translates into problems of access to justice for citizens and businesses, such as a lack of regulatory and legal certainty and a deterioration of the business environment. Although it is generally accepted that human decision-making by judges should not be *replaced* **by AI, specific use cases are gradually becoming widespread for assisting judges in their duties.** AI can be of assistance in many ways, as shown in Table 4 below by the Council of Bars and Law Societies of Europe (CCBE) (2020). In general, the CCBE makes it clear that judges shall not be replaced in their decision-making. AI may be deployed to assist and *possible* uses of AI range from case management to pre-trial and in-trial applications, as well as tools that aim to support judges in deliberation and decision-making phases and so-called 'post-sentencing applications', making AI deployment gradually more widespread in this domain (Ronsin and Lampos, 2018, p.42).

Considering the capacity to gather extensive evidence through data and AI, Pagallo and Quattrocolo (2018) discuss whether the use of investigative intrusions through AI violates the right to private life and the right to a fair trial. The lack of 'fair balance' between parties is likely to occur if automated evidence gathering does not allow for transparency of how the data was gathered (e.g. through deep neural networks). The authors conclude that the right to private life (Article 8 EHCR) and the right to a fair trial can be seriously affected if the evidence is collected and processed with non-transparent self-learning machine algorithms.

Table 3 - Uses of AI by courts

Stages	Management of cases	Pre-trial	Trial	Judges' deliberation/ decision-making	Post sentencing
(Potential) AI applications	- Case management system - Electronic communications - Digital platforms accessible for lawyers/clients - Automatic monitoring of procedures - Automatic system for monitoring procedural delays - Automatic system for completing procedural formalities - Automatic decisions on the progress of the case - Queue management - Automatic sorting of appeals	- Plea-bargaining: Prosecutor's databases	- Use of videoconference - Automated transcription / automated translation - Automated presentation of file's document on screens during hearings - Case management (in a situation of complex cases) - Use of emotional AI (detection of emotions, etc....)	- Case law tools - Prediction technology - Legal researches and analysis / autonomous researches - Scoring of risks / assessment of the suspect (probability of recidivism) - Automated judgments (decision trees) - Writing assistance tools and drafting judgments - Decision making systems - Intelligence assistant systems (identification of patterns, analysis of data...)	- Scoring of risks / probability of recidivism / parole opportunities
Main principles and issues to be taken into account					
Principles	- Adversarial proceedings - Rule of law, due process, security - No restriction of access to justice - Equality of arms - Transparency of decision-making - Access to data by lawyers	- Adversarial proceedings - Equality of arms - Access to data by lawyers - Data protection and compatibility with fundamental rights	- Adversarial proceedings - Fair trial - Transparency - Neutrality (in profiling) - No use of emotional AI when videos are used during a trial	- Adversarial proceedings - Fair trial - Transparency about use of AI by judge - Transparency of decision-making process - Algorithms and accountability - Liability if errors occur - Access to evidence - Right to request for a human intervention (judge)	- Adversarial proceedings - Fair trial - Transparency of decision-making process - Algorithms and accountability - Right to appeal

Source: CCBE (2020)

In their submission to the consultation on the EU White Paper (see Chapter 2), Guild et al (2020) argue that 'AI, if unregulated or regulated ineffectively, may lead to the breach of fundamental rights, including the rights to an effective legal remedy and a fair trial, as protected within the EU by Article 47 of the Charter, Article 6 ECHR and the general principles of EU law.' The particular challenges involved in ensuring access to a remedy and procedural

fairness from ADM relate to transparency, unpredictability and complexity, which run directly contrary to the rule of law.

In its submission to the White Paper on AI, the CCBE (2020) describes the many principles that might be impacted by the use of AI tools:

- Use of data and elements that have not been the subject of an adversarial debate.
- Exploitation of conclusions (even partial) that have not been obtained through the reasoning of the judge.
- Lack of transparency of the process, since it becomes impossible to know what should be attributed to the judge and what comes from a machine.
- Absence of a level playing field (equality of arms).
- Undermine the principle of impartiality due to the impossibility of neutralising and knowing the biases of system designers.
- Breach of the principle of explicability due to the existence of results that are beyond human reasoning and cannot be traced.

Potential bias in the datasets used to train an AI model clearly affects the fairness of a trial. Many AI systems function on statistical correlations without any human understanding of societal contexts. Input data is the only context in which AI systems operate and if the data provided to train an AI model or as its input is incomplete or include (even non-intentional) problematic bias, then the output of AI can be expected to be incomplete and biased as well. AI systems still lack transparency (CCBE, 2020) and 'explainability' (the ability to explain both the technical processes of an AI system and the related outputs). Bias could be harmless in most situations but could also be harmful especially when AI systems are used before a court that conclusions based on them may be insufficiently substantiated to ensure the fairness of the proceedings.

Likewise, the use of AI in criminal justice can lead to inherent bias in predicting crime or assessing the risk of re-offending, while facial recognition technology is known to be inaccurate at identifying people of particular ethnicities[43]. In the US, the Electronic Privacy Information Centre (EPIC) offers a detailed overview of the risk modelling tools currently in use by public administrations (EPIC, 2020). Discrimination based on ethnicity poses a threat to civil rights. Moreover, defendants would likely be unable to challenge predictions made by algorithms because the decision-making process of the algorithms is not disclosed. Applications of AI in forensic work and re-offence risk assessment are thus problematic. Another concern relates to the inequality between the more advanced capabilities prosecutors possess and the relatively limited resources of lawyers (CCBE, 2020). Ontier (2020) observes that no court in the EU Member States is using predictive technology solutions to make judgments based on AI software, unlike the US, where AI is already in use (e.g. the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) (Ortiz

[43] The EU Commission for the Efficiency of Justice (CEPEJ) developed the European Ethical Charter on the use of Artificial Intelligence in judicial systems and their environment, setting out guidelines on automated processing of decisions and judicial data based on AI. This document establishes five principles that must be looked at in order to develop AI tools to be applied to the judicial system: (i) Principle of respect for fundamental rights: design and AI services must not infringe fundamental rights; (ii) Principle of non-discrimination: any discrimination between individuals or groups of individuals must be avoided and prevented; (iii) Principle of quality and security: with regard to the processing of legal files, decisions and data, relating to using certified, reliable sources and always working within a secure technological framework; (iv) Principle of transparency, impartiality and fairness: processing of data must be made in accordance with the principle of transparency and external audits must be performed; (v) Principle 'under user control': aiming to ensure that users are properly informed and have control over their actions.

Hernández et al., n.d.)[44]. In a landmark decision, the Supreme Court of Wisconsin decided in *State of Wisconsin v. Loomis* (2016) that 'a circuit court must explain the factors in addition to a COMPAS (a system based on an algorithm) risk assessment that independently supports the sentence imposed. A COMPAS risk assessment is only one of many factors that may be considered and weighed at sentencing' (p.49, para 99).

Several other uses of AI in government have raised concerns among scholars and activists, with examples from various parts of the world. In addition to the risk to the right to a fair trial, concerns are expressed about the risk to the principle of presumption of innocence, defence rights (such as the right to be informed about the use of AI) and the right to an effective legal remedy where AI is used. In a report for the EPRS, Gonzalez Fuster (2020) notes that the use of AI by governments is already a reality. This raises concerns mostly in the fields of 'predictive policing, facial recognition, AI and criminal justice, and AI and borders (including a reflection on the European Travel Information and Authorisation System, ETIAS), for instance in litigation and calls from civil society to better prevent or mitigate associated risks, both in the EU and beyond'. She also notes that the current EU data protection legal framework 'shall not be assumed to offer enough solid safeguards for individuals' in light of the increased uses of ADM and profiling for law enforcement and criminal justice purposes. This is critical since the general safeguards provided by the GDPR do not necessarily apply when the processing is for such purposes, as restrictions and derogations might be applicable. While the Law Enforcement Data Protection Directive is not exactly equivalent to GDPR and also provides for possible restrictions and derogations, some EU safeguards do pertain, as the Directives provides for the right not to be subject to ADM.

In its submission to the consultation on the AI White Paper, the German Bar Association's Committee on European Affairs provided several examples of potential violations of fundamental rights through the use of AI in justice and enforcement. In China, the so-called 'cyber-court' transferred the entire administrative procedure for case handling online. Since 2019, the Supreme People's Court operates a 'mobile court' pilot programme, in which an AI-driven chatbot 'judge' manages civil procedures through the country's social media platform WeChat and the evidence is entered into a blockchain. Other cases in which AI assists judges include the *Prédictice*[45] system tested in 2017 by the courts of appeal in Rennes and Douai, now in use in law firms, alongside similar products (Luminance, Nakhoda, Kyra System, etc.). *Prédictice* uses open data provided by the French government and text analysis on law cases to support law professionals to analyse and evaluate cases.

Some predictive policing tools are deployed during the post-sentencing phase, and a general distinction can be drawn between pre-trial risk assessment tools and risk assessment tools to evaluate re-offending in the decision-making phase. In some cases, individuals did not commit a crime but are subject to risk assessment because of their contact with criminals or their social background. Applications similar to the most prominent US AI system (COMPAS) are emerging in EU Member States' law enforcement systems, such as the *ProKid* AI tool that was used in the **Netherlands**. ProKid aimed to identify the risk of recidivism[46] among 12-year old children previously been suspected of a criminal offence by the police. A similar tool (*SAVRY*) is used by Spanish authorities. However, the interplay of such AI-supported risk

[44] See also PROMETEA Software of Artificial Intelligence aimed at streamlining and optimising bureaucratic processes in all types of organisations, developed by the Public Prosecutor's Office of the City of Buenos Aires and the Law School of the University of Buenos Aires (https://ialab.com.ar/prometea/).

[45] For more on *Prédictice,* see https://predictice.com

[46] Prokid is not strictly a post-sentencing tool as children have not been convicted, but it is used to predict the potential risk of offending.

assessment tools with the principle of presumption of innocence is still insufficiently addressed.

Further legal research tools are another practical example of AI used in the domain of justice. The **Italian** programme *TOGA*, for instance, serves as a database for prosecutors (and lawyers). Lawyers and insurers are increasingly relying on AI-tools, especially those that predict a judge's decision. A typical example is *Jurimetria*, a statistical and predictive jurisprudential software that helps legal professionals in Spain to analyse their cases. It systemises and extracts content from more than 10 million judicial decisions from all instances and jurisdictional orders in **Spain**. Another prominent example is *Casecruncher Alpha*, which, in October 2017, won a week-long competition against human commercial lawyers with a prediction accuracy rate of 86.6%. At first glance, predictive analytical tools used by lawyers do not appear to hinder access to justice. However, it is important to keep in mind that the work of lawyers goes beyond providing a brief legal response to a simple question.

The use of AI in courts' administrative systems could affect fundamental rights if used in a targeted manner. Such concerns became real when the Ministry of Justice in **Poland** introduced a system of algorithm-driven allegedly random allocation of cases. The digital system assigns cases to particular judges across the country on a once-per-day basis. If the system were truly random and left no discretion to its operator, this would not appear problematic at first sight. It was argued, however, that the Prosecutor General could unduly influence the process. As part of the Ministry of Justice and thus a party to criminal proceedings, the Prosecutor General could control how cases would be assigned. Such influence could ultimately result in a violation of the right to a fair trial. The concerns in this example were aggravated by the fact that the Ministry was unwilling to disclose the workings of the algorithm used for the system.

The rule of law might be further endangered by the use of AI tools in law enforcement. Such tools (e.g. Prédictice) could be applied directly in the courtroom or play an indirect role as a basis for a decision challenged in a court proceeding. The issues here are that affected individuals are not usually aware that these tools are being used, and police may not wish to publicly disclose the criteria that determine the AI system's outcome, nor how the criteria are weighed nor the data used to train the system's algorithms. Such systems prevent access to justice, as the affected individuals can neither detect nor prove whether they have been subject to an erroneous or unfair decision. The systems collect considerable amounts of data, which may be hacked and lead to grave data protection and privacy infringements. One particularly important example is the EU-funded iBorderCtrl-research project (Intelligent Border Control System)(Leufer and Jansen, 2020), which tests software that aims to detect when people are lying at border controls: third-country nationals are asked to answer questions from a computer-animated border guard avatar, which analyses their micro-gestures to decide if they are lying. According to an analysis by AlgorithmWatch, the system contained a strong risk of racial bias, as it was mostly trained on white European men and had a high error rate (around 25%).

Looking ahead, one of the priorities of the EU's 2019-2023 e-Justice Action Plan is to take stock of the use of AI, blockchain and distributed ledger technology (DLT) deployed in the justice systems. The European Commission has released a study on the use of innovative technologies, which identified 130 projects/uses cases of AI and blockchain technologies deployed within judicial processes (Vucheva et al., 2020). The study also identified eight issues which were then mapped against eight business solution categories. The report acknowledges that closer cooperation is needed at the EU level to exchange good practices, avoid duplication of effort among the Member States, and explore synergies, and suggests some actions and mechanisms to strengthen that cooperation.

Good administration: transparency and accountability

Every individual enjoys the right to good administration, under Article 41 of the EU Charter. This includes the 'right to have his or her affairs handled impartially, fairly and within a reasonable time by the institutions, bodies, offices and agencies of the Union', 'to be heard, before any individual measure which would affect him or her adversely is taken', and 'to have access to his or her file, while respecting the legitimate interests of confidentiality and of professional and business secrecy'.[47] Importantly, this right translates into **an obligation for administrations to give reasons for their decisions.** Finck (2020) discusses the issue of transparency as an element of the right to good administration, with specific respect to ADM. AI implementation can significantly affect the enjoyment of this right in at least two ways: by potentially discriminating between citizens through social scoring systems, inferential analytics, and machine learning applications that inevitably carry a risk of discrimination, profiling and intrusion in citizen's private sphere, resulting in a lack of equal access to public services; and when governments use machine learning systems that are not interpretable and explainable, depriving citizens of the right to an adequate explanation for decisions adopted by the administration. Amsterdam and Helsinki are to launch open AI registers that track how algorithms are used in their municipalities in order to increase the principles of responsibility, transparency and security in the use of AI in public administration (Macaulay, 2020).

According to a recent study for the Administrative Conference of the United States (Engstrom et al., 2020), US administrative agencies have already used various AI tools across different government tasks, including law enforcement, single-case decision-making, monitoring and analysing risks to public health and safety or other policy objectives. They also apply them to extract information from the government's data resources communicate with citizens and business, and perform intra-administrative management of resources, including procurement and maintenance of public facilities.

Migration policy and AI

AI deployment has become extremely widespread and controversial in the domain of border controls, and migration policy generally. Gonzalez Fuster (2020) and the EDPS (2017) observe that *eu-LISA* (EU Agency for operational management of large-scale IT systems in the EU), upcoming EU-wide information systems including the Entry/Exit System (EES), the European Travel Information and Authorisation System (ETIAS) and the European Criminal Records Information System for Third-Country Nationals (ECRIS-TCN) increasingly deploy AI. Likewise, the revised Schengen Information System (SIS) has announced to use facial recognition technology, DNA and biometric data[48]. The collection and use of data through AI systems may lead to significant violations of fundamental rights, such as non-discrimination and the right to good administration: existing trials include *iBorderCTRL* (see above). In the US, similar systems such as *SilentTalker*, *EyeDetect* and *Discern* are being trialled privately or even by public administrations, on the assumption that lying is more cognitively demanding than telling the truth (see the corresponding section). Beduschi (2020) reports that Canada already deploys algorithmic decision-making and AI technologies for immigration and asylum processes (Molnar and Gill, 2018). Likewise, Switzerland is experimenting with an algorithm

[47] See https://fra.europa.eu/en/eu-charter/article/41-right-good-administration#:~:text=Article%20XXIV%20(Freedom%20and%20Responsibility,provided%20for%20by%20an%20Act.

[48] One example of system being used at the border is the 'Passage automatisé rapide des frontières extérieures' (PARAFE), based on the automated control of biometric passports, either through analysis of fingerprints or facial recognition technologies.

for the integration of refugees. However, as a downside, Beduschi underlines growing reservations about the emergence of a form of 'surveillance humanitarianism' (Latonero, 2019).

c. Other fundamental rights affected by AI

This survey of the risks created by current and emerging uses of AI for fundamental rights has focused on the specific aspects that are most evident in current research, including discrimination, human agency, freedom of expression and privacy. However, the features of AI systems described above mean that AI may impinge on other fundamental rights. These are described briefly below.

- The deployment of AI solutions in the B2C context has far-reaching consequences for **consumer protection**. The widely researched informational asymmetries that characterise consumer markets are amplified by the use of AI tools aimed at enhanced profiling, price differentiation, hyper-nudging and collection or inference of tests, interests and consumers' willingness/ability to pay. All of these tools also provide for potentially welfare-enhancing market practices, such as the ability of firms to customise their conditions and product offering to perfectly match consumer taste, the elimination of cross-subsidisation through efficient price discrimination, and even (in the IoT age) the drastic reduction of transaction costs through the use of automated transactions (e.g. the dash replenishment button used by Amazon). However, where AI systems are not fully explainable and interpretable, their use can dramatically reduce consumers' ability to interact, compare conditions they are awarded market indicators, gauge the level of discrimination to they are subject, and, in certain circumstances, even discern the actual price/product. The literature acknowledges that in complex multi-sided platforms, prices of certain services tend to reach zero or even negative values, as consumers offer (often inadvertently) their data and attention to businesses, including platforms, online intermediaries and advertisers. Features of AI systems earlier in this report (profiling, hyper-nudging, echo chambers, data aggregation, inferential analytics, etc.) reverberate on consumers, creating a number of new risks and exacerbating well-known imbalances of the B2C environment.
- The **right to freedom of assembly and association** is particularly affected by the use of AI tools to identify participants in public gatherings, including demonstrations. Here, a tension may emerge between the need to protect public safety and security, and the protection of individual fundamental rights.
- **Sustainability and protection against sustained impairment of the living standards of future generations** by intelligent systems (Hoffmann-Riehm in Wischmeyer & Rademacher, 2020) can be interpreted as aspects related to the EU Charter provision on environmental protection (Article 37). With the EU's commitment to the Sustainable Development Goals (SDGs) and the principles formulated in the European Green Deal[49], it is worth mentioning the negative impact of AI systems, due to high amounts of energy consumption and substantial amounts of technological waste. A recent MIT study found that 'training one AI model produces 300,000 kilograms of carbon dioxide emissions, roughly the equivalent of 125 round trip flights from New York to Beijing' (Strubell et al., 2019). This is because the computational resources needed to improve the accuracy of machine learning models require

[49] The European Green Deal 'is a new growth strategy that aims to transform the EU into a fair and prosperous society, with a modern, resource-efficient and competitive economy where there are no net emissions of greenhouse gases in 2050 and where economic growth is decoupled from resource use' (Communication/2019/640 final, p. 1).

substantial energy consumption, making AI 'costly to train and develop, both financially, due to the cost of hardware and electricity or cloud compute time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware.' Thus, large-scale AI deployment is a raw material-intense endeavour and the impact of an expected AI uptick in the near future are insufficiently researched.

More generally, AI systems can lead to discrimination due to lack of consideration for disability and whenever their widespread diffusion puts people with lack of digital skills at a disadvantage. The European Disability Forum voices specific concerns that the needs of disabled users are not sufficiently taken into account in the design and deployment of AI systems (Marzin, 2018).

2. AI risks for safety and security: a systematic literature review

AI 'refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals'[50]. AI is a technology that offers revolutionary and positive developments, ranging from leisure (video games), to manufacturing, finance and even government/military use. However, the benefits also carry risks, whereby the AI itself may pose threats to digital, physical and political security.

To identify the major AI safety and security risks, the study team undertook a systematic literature review (SLR) on selected academic journals, reliable sectoral magazines, websites and white papers, official documents from governments and international organisations, position papers of civil society organisations, and official government decisions and case-law.

a. Safety risks caused by AI

AI advancements offer great opportunities in cross-cutting realms impacting all layers of society, such as health, business, or education. However, these opportunities also entail major risks to safety and security. Identifying safety and security risks caused by AI thus appears to be an essential task for the EU and its Member States if they are to build 'trustworthy' (i.e. lawful, responsible, sustainable and safe) AI, allowing the full potential of this technology to be harnessed, while mitigating its negative aspects. As Renda (2019) puts it: 'As we take our first steps in this blossoming new world, we can still decide how AI can help us promote a better society and a more sustainable future. In other words, we have the chance to approach policy choices in the best possible way: by asking the right questions, at the right time and in the right sequence.' These rights questions entail close scrutiny of safety and security issues.

The SLR identified key elements related to the potential risks caused by AI. The notion of 'risk' is defined as a situation relative to a danger caused directly or indirectly by the development and/or deployment of AI. As explained by Yampolskiy (2016), AI safety is linked in the literature to the concept of 'Safe AI'. The notion of AI safety also arose from debates on ethics in AI, especially fundamental questions of long-term risk and impact on human society. Researchers have explored the necessary legislation and product liability of AI in respect of potential failures that impact different segments of society.

[50] EC (2019) "Trustworthy AI – Brochure" https://ec.europa.eu/digital-single-market/en/news/trustworthy-ai-brochure

Context and background in the EU

The private and public sectors, as well as citizens, benefit from the use of AI as an emerging technology. Nevertheless, there are inherent risks to rights, legal certainty and safety. Citizens in particular may face unintended effects of AI that can be used for malicious purposes. The European Commission published a political communication setting out seven key requirements that AI applications should respect to be considered trustworthy (European Commission, 2019).

The White Paper on AI by the European Commission (2020a) groups AI harm into two categories:

- Material risk: safety and health of individuals (including loss of life) or damage to property.
- Immaterial risk: loss of privacy, limitations to the right of freedom of expression, human dignity, discrimination (e.g. in access to employment).

The report from the Expert Group on Liability and New Technologies on Liability for Artificial Intelligence (2019) points out that, in the EU, product safety regulations ensure that new technologies minimise the risk of bodily injuries and harm. However, regulations do not 'completely exclude the possibility of damage resulting from the operation of these technologies' (European Group on Ethics in Science and New Technologies, 2019).

When embedded in products and services, AI technologies may present safety risks. For example, a flaw in object recognition technology may lead an autonomous car to wrongly identify an object on the road and cause an accident. These risks may be caused by the design of AI systems, or any issues related to the availability and quality of data input into a machine-learning process. AI safety risks are also highly linked to legal certainty, where if the requirements are not met, European companies' competitiveness may be undermined. The embedding of AI systems in a product or service might make it difficult for a person who has suffered damage to retrace back the fault to the AI technology.

The European Commission's Report on the safety and liability implications of artificial intelligence, the Internet of Things and robotics (European Commission, 2020b), accompanying the White Paper, analyses the relevant legal framework with respect to specific risks posed by AI systems and other digital technologies. The following safety risks are highlighted:

- Autonomous behaviour of certain AI systems;
- Mental safety risks (e.g. collaboration with humanoid robots);
- Faulty data at the design stage and maintenance of data quality throughout the use of AI products and systems;
- Opacity of systems based on algorithms;
- Impact of stand-alone software;
- Complexity of supply chains.

The current product safety legislation of the EU supports a large number of risks arising from a product itself, but the risks specific to AI necessitate further legal certainty before their full trustworthy use. The current EU safety framework consists of sector-specific legislation, such as the Machinery Directive (Directive 2006/42/EC), the Radio Equipment Directive (Directive 2014/53/EU (RED)), the Measuring Instruments Directive (Directive 2014/32/EU), and the

General Product Safety Directive (Directive 2001/95/EC), a horizontal instrument establishing general safety requirements for consumer products in the EU. While these pieces of legislation were adopted prior to the emergence of AI, the Machinery Directive and the General Product Safety Directive are among those currently under review for adaptation to new technologies, including AI.

Taxonomy of AI safety risks

One starting point to ensure AI safety lies in the data source (input and training data), seeking to ensure that the AI system is built on safety from the earliest possible stage.

Scott and Yampolskiy (2019) develop an AI incident taxonomy, building on Hollnagel (2014) and Yampolskiy (2016) and deconstructing safety into **consequences** (phenomenology), **agency** (aetiology), **preventability** (ontology), and **stage of introduction in the product lifecycle** (phenomenology and ontology), all complemented by further literature.

Consequences

- Consequences of AI safety failures impact individuals, corporations and communities, as reported by Scott and Yampolskiy (2019). These consequences are:
- **Physical**: individuals may face harm at different levels, ranging from inconvenience to loss of life.
- **Mental**: individuals' mental health may be impacted by new beliefs that were propagated through fake news or chatbots, among others.
- **Emotional**: individuals may suffer mental consequences, given AI's new roles in society, leading to the possibility of depression.
- **Financial**: individuals, corporations, and communities all face financial consequences, positive and negative, from AI uptake.
- **Social**: AI can lead to the modification of individuals' behaviour.
- **Cultural**: AI can lead to modifications of individuals' vision, values, norms, systems, symbols, language, assumptions, beliefs, and habits (Needle, 2004 cited in Scott and Yampolskiy, 2019).

In today's society, all of these aspects are interlinked. A beneficial financial consequence to a corporation of replacing a task with AI instead of humans may negatively impact individuals' mental health.

Agency

'The agency of a failure is the degree of human intentionality in its origin or propagation' (Scott and Yampolskiy, 2019, p. 4). The authors classify these agencies as accidental, negligent, intentional, and malicious. Safety is associated with accidental risk, and security with malicious intent.

AI safety risks arise from threats from a machine learning system. According to Amodei et al. (2016), accidents are defined as unintended but harmful behaviour that may emerge from the poor design of certain AI technologies or systems. The authors identified five possible failure modes and concrete problems in AI safety, as depicted in the image and further described below.

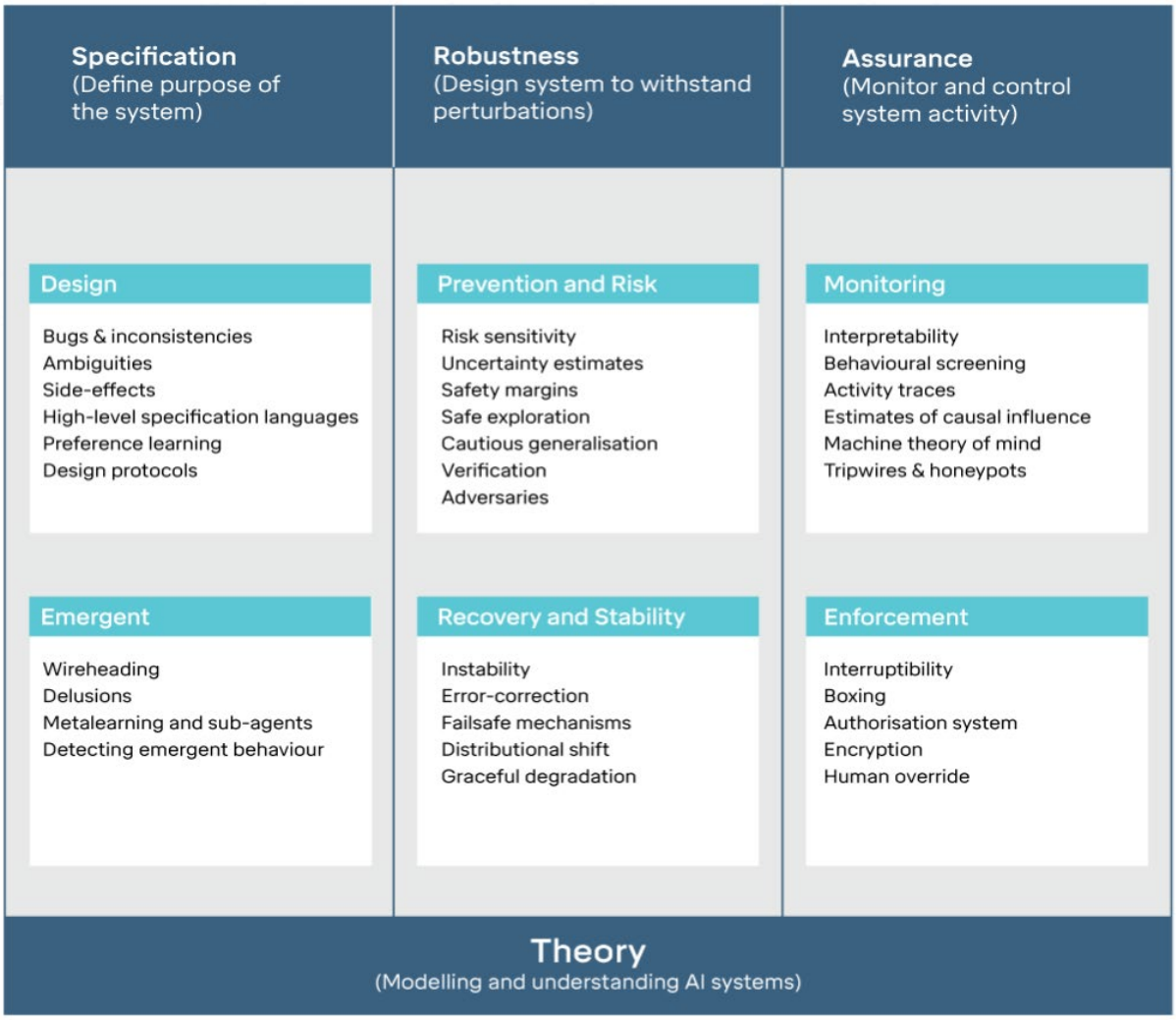
Figure 4 - Classes of AI safety issues

Source: Amodei et al. (2016)

- **Safe exploration**: an autonomous agent needs to engage in exploration, i.e. 'taking actions that don't seem ideal given current information, but which help the agent learn about its environment' (Amodei et al.,2016, p.14). Although these situations can be potentially dangerous in the chosen environment, hard coding offers the possibility to avoid catastrophic behaviours.
- **Distributional shift**: an AI system relies on its testing distribution to perform in the real-world/training distribution, where any factors with which it is unfamiliar may cause poor performance, without the system understanding that its action was wrong.
- **Negative side effects**: objective function to focus on a single aspect of the environment and overlooking the rest, causing disruptions.
- **Reward hacking**: a system discovers possibilities to gain a reward by not completing the exact task at hand, for example creating a new problem to solve it or ignoring and not reporting the problem at hand.
- **Scalable oversight** or semi-supervised reinforcement learning (Christiano, 2016): a system can be tasked with multiple steps that offer a reward ahead of the final evaluation.

Ortega et al. (2018) define three areas of technical AI safety: specification, robustness, and assurance, as depicted in Figure 6.

Figure 5 - AI safety categorisation



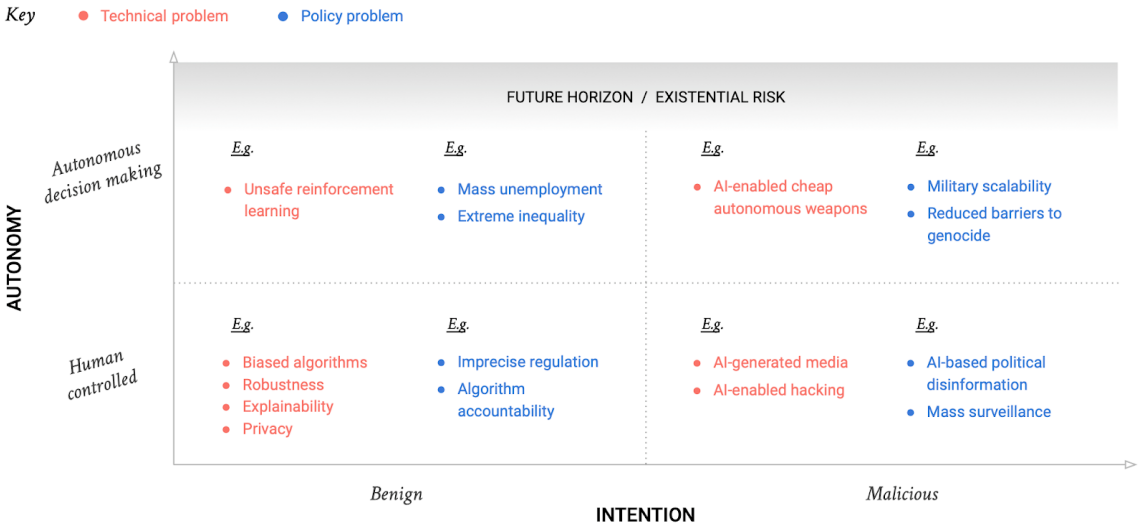
Source: Ortega et al. (2018)

Ortega et al. (2018) highlight the challenges and approaches of the three categories:

- Specification ensures that an AI system's behaviour aligns with the operator's true intentions.
- Robustness ensures that an AI system continues to operate within safe limits upon perturbation.
- Assurance ensures that we can understand and control AI systems during operation.

Feige (2019) categorises agency according to benign and malicious, as depicted in Figure 7.

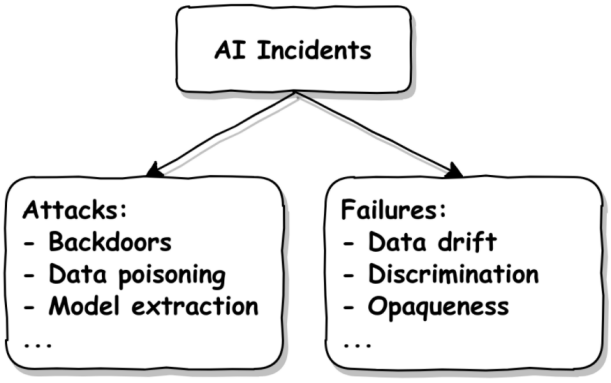
Figure 6 - AI safety space



Source: Feige (2019).

The basic taxonomy of AI incidents is explained by Burt and Hall (2020). Figure 8 below shows their two categories: (i) attacks and (ii) failures.

Figure 7 - AI incident taxonomy



Source: Burt and et P. Hall (2020)

The authors define an AI incident as '*any* behaviour by the model with the potential to cause harm, expected or not'. This taxonomy therefore divides AI incidents into malicious attacks and failures.

This literature review focuses on safety as depicted above. However, the classification below is also important in academic research on security and malicious intent. Specific AI incident attacks are:

- **Backdoor risk:** unauthorised access to a system to create a malicious trigger, as the attacker must insert it during the system's training phase. The peculiarity of backdoors

in AI models is that a machine learning algorithm such as a neural network corresponds to an aggregation of different parameters aiming to interact with the data. Thus, the fact that there is no 'source code' per se, as in traditional programming, makes the detection of backdoors far more complex in AI and in the machine learning model supply chain (Gu et al., 2017). In other words, the millions of parameters of an AI system cannot be inspected in the way that traditional code might be.

- **Data poisoning:** an attack directly targeting the data of a model (European Organisation for Security, 2019). An example of data poisoning is Microsoft Tay, a chatbot supposed to interact with young people on social media that was flooded with offensive and racists tweets in 2016 (Wavestone, 2019). This data poisoning resulted in the subversion of the bot's initial use so that it started publishing inappropriate content on Twitter. Data poisoning can affect a vast array of datasets, such as healthcare data, loan or house pricing (Alfeld, 2016).
- **Model extraction:** when the attacker tries to extract parts of different classes from a machine learning model (classes on which the model was trained). Such attacks not only represent an intellectual property issue but can also lead to certain dangers. Carlini et al. (2019) managed to use this method to extract credit card numbers and social security credentials to expose the safety and security threats that could arise from model extraction.

Failures can then become safety risks:

- **Data drift**: Data scientists must monitor model performance over time to ensure that the AI system is behaving and learning as necessary. Chowdhury et al. (2020) explain, 'Over time, a machine learning model starts to lose its predictive power, a concept known as model drift. What is generally called data drift is a change in the distribution of data used in a predictive task.' The underlying functions of the system start changing and the model accuracy degrades over time, which may lead to risks if not properly controlled, monitored, and tested.
- **Opaqueness**: The black-box feature of AI and machine learning is referred to as opacity, which is a major concern in guaranteeing the safe exploitation of AI products on the market. Quality assurance to identify bugs is difficult in AI and is approximative work, as there is a reliance on the data and algorithms (Schmelzer, 2020). The opacity of systems may cause material damage to property or physical harm to users, as retracing the steps of an algorithm may be impossible.
- **Misdirected reinforcement learning behaviour**: This is a risk inherent to the functioning of certain types of AI applications. Where an application is set to learn from trial and error, some of the 'trials' can lead to undesired consequences if the system is not appropriately constrained in its outputs.

The complexity and key takeaway reside in AI failures, which are difficult to assess, as they 'can be caused by accidents, negligence or unforeseeable external circumstances' (Burt and Hall, 2020). Accidental factors refer to the possibility of a negative and dangerous outcome stemming from an AI model or AI use. The main issue with the accidental factor is the generation of unintended risks because of the inability of the model to comprehend its environment properly. One of the most illustrative examples is the self-driven car, whose collision with other cars or humans could result in fatality.

Degree of preventability

Scott and Yampolskiy (2019) describe four degrees of AI failure preventability:

- Trivially preventable;
- Preventable with some difficulty;
- Preventable with great difficulty;
- Unpreventable.

The foreseeability of systems is a core concept in the realm of product safety. If a producer is unable to ensure the safety of a product incorporating AI because the functionality (e.g. self-learning) makes certain product features unpredictable, the product should not be released to the market. In addition, the outputs of AI systems are only predictable to a certain extent - some may become so powerful that any of the failures depicted may become unpreventable, giving the system a strategic advantage over human control.

Product lifecycle stage

Depending on the phase of the AI product, the risks tend to differ significantly. AI developers must assess these challenges to ensure the safety and health of society before placing the system on the market. The sources studied reveal a distinction in risk typology between (i) **the development phase** of an AI project, (ii) **the deployment phase** of an AI project (also called the *learning phase,* where the AI is built and the *processing phase* where the AI is launched, see Wavestone (2019)), and (iii) the **adaptation phase** of an AI project.

Following McKinsey's (2019) mapping of 'the different risks spanning the entire life of an AI solution', there are three steps in which specific types of risks can arise during the ***development phase*:**

- Conceptualisation;
- Data collection process;
- Model development phase.

During the ***deployment phase***, two steps with corresponding risks can be identified:

- Model implementation;
- Model use and decision-making phase.

Scott and Yampolskiy (2019), along with Dawson, Burrell, Rahim, and Brewster (2010) claim that 'the cost of fixing an error at each stage is ten times the cost of fixing it in the previous stage' (p.5).

To put these phases into perspective in the healthcare sector, Figure 9 shows the AI translation workflow during the development, deployment and adaptation phase of an AI system (Hu et al. 2020).

Figure 8 - AI translation workflow

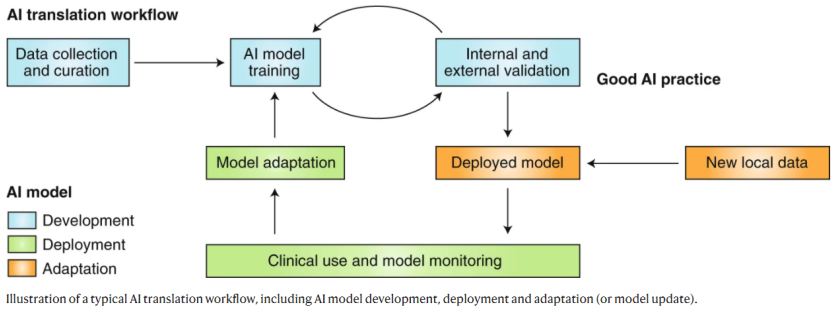


Illustration of a typical AI translation workflow, including AI model development, deployment and adaptation (or model update).

Source: Hu et al. (2020)

Hu et al. (2020) argue that 'rigorous validation is key to ensuring that safety and efficacy are tested; models must be validated before initial deployment and continuously monitored and adapted when implemented in local healthcare environments and as outcome likelihoods change due to evolving patient management strategies'.

Industry and sectoral overview of safety risks

AI can cause safety risks in new products, as well as where it is integrated or embedded into an existing product. In the latter case, unforeseen safety problems may arise due to system dependencies, interactions with other products, data incorporation, interactions with the environment. AI can be dangerous or risky for safety in different ways: it can make the wrong decisions, it can forego context, empathy and emotions, and it may introduce bias in decisions. This section examines an industrial and sectoral approach to identify safety risks that arise due to AI. The sectors covered are healthcare, transport, energy, and the public sector, as depicted in the Commission's White Paper.

Healthcare

AI cannot be deployed in healthcare without sufficient infrastructure and risk mitigation. Failures in the medicines industry may cause harm to individuals by, for example, failed robotic surgeries or incorrect treatment (Ross and Swetlitz, 2018). 'Bias in the operation of an algorithm recommending specific treatment could create real health risks to certain groups' (OECD, 2019). At times, AI systems are also associated with low prediction accuracy (Ellahham et al., 2019).

Macrae (2019) points out the potential of AI systems in healthcare, for instance, to increase diagnostic accuracy and optimise treatment planning, or even to forecast outcomes. However, replacing the human knowledge and 'subjective' analysis of a patient carries numerous safety risks. One failure in the system – whether due to hidden assumptions, incorrect recommendations, or errors in analysis of an image (e.g. tumours, moles, etc.) - may impact hundreds of patients. IBM's Watson supercomputer launched a medical AI recommendation system for patients with cancer in 2018, but the system made inaccurate treatment recommendations. Human oversight proved lifesaving, as some recommendations could have been fatal (Ross and Swetlitz, 2018)[51].

[51] https://mc.ai/what-is-the-reason-of-ibm-watsons-failure-in-healthcare/

The WHO reports only limited cases of transferring decision-making capacity to AI systems as yet. Human oversight is still necessary for performing interventions, and software supported by AI systems has to undergo strict testing (e.g. controlled infusion pumps) (Habli et al., 2020). Building trust in the system among both practitioners and patients is essential to ensure acceptance of the safety risks caused by AI systems. Collaboration between healthcare practitioners and (health) data scientists will lead to a better understanding of the safety risks that may emerge in the field. The European Patent Office has added guidance for patent applications for AI-based and ML devices (2019).

Transport

In transportation, 'AI uses observed data to make or even predict decisions appropriately' (Antony, 2017). Safety risks may arise, however, because transport is a sector where unpredictable actions in traffic can lead to accidents. For example, pedestrians and cyclists may move in unforeseen manners, and an AI system, without human (sense) oversight, may not detect a wrong movement. AI can aid to overcome human error, but it can also lead to safety risks for pedestrians, passengers and others in the environment.

Intelligent Transport Systems (ITS), autonomous vehicles (AV) and AI complement each other. In both urban and motorway settings, vehicles are being developed to include 'smart devices' (e.g. lane keeping, cruise control, anti-collision braking). The sensor data gathered by vehicles is shared with manufacturers, who then use the data to improve software (possibly including the training or retraining of machine learning models). No significant learning takes place in the vehicles. Nevertheless, as more automatic driving tasks are performed by AVs, there is a strong possibility of long-term deterioration in (human) driver performance, thereby causing potential road safety risks (Miles and Walker, 2006). This may be due to drivers trusting and relying on the various devices in their car[52].

The most important safety risks in transport and AI are due to system failures, which may be critical. When a human-based task is replaced with AI in transport, citizens face safety risks, such as self-driving vehicles and trucks without human interaction that rely fully on AI for safety. Although Uber and Tesla have developed self-drive cars and trucks that they claim will reduce the number of accidents, semi-AVs killed a pedestrian in the case of Uber (Lee, 2018) and the driver in the Tesla case (The Guardian, 2018) because of the system's self-override function.

Smith (2020) argues that human driving may pose higher safety risks than automated driving. Smith notes that failures such as those above can be detected and fixed. The underlying question is how safe a vehicle should be for deployment, taking into consideration a potentially dangerous automated driving system that may cause damage (for which someone would be liable). Especially in transport, 'safety is an ongoing process that begins before product development and continues through product disposal'. A safety conformity assessment is necessary in both a 'protected' environment and openly on the road in order to interact with other road users, learn weather and road conditions, recognise obstacles. It is in all of these environments that AI learns and develops its data (Niestadt, 2019).

Energy

[52] The Society of Automotive Engineers (SAE) defines six levels of driving automation, ranging from 0 (fully manual) to 5 (full automation). Current commercial cars have reached levels 2-3, where the human monitors the driving environment and some vehicles can perform most driving tasks.

The adoption of AI in the energy sector is slow compared to other industries such as education, healthcare and transport, and its safety implications are not (yet) a popular research topic. Top uses of AI deployment within the energy sector include energy monitoring and management, wind power analysis AI platforms (by Google and IBM[53]), and wildfire powerline and gear monitoring (e.g. wildfire preventability by AI-powered early detection systems). As noted at the UN AI for Good Global Summit in 2019, AI systems can address climate change.

Energy systems are critical infrastructures whose data make them susceptible to terrorist attacks. Such cyber vulnerability was evident in 2015 and 2016 in Ukraine, when a power grid was attacked, leaving thousands of people without power (Cerulus, 2019). Currently, governments (including the US Pentagon's Defence Advanced Research Projects Agency) are investing more into researching security vulnerabilities rather than safety risks caused by AI. In the energy sector, developments do not rely solely on AI - thorough analytics, sensors, robotics and IoT devices are necessary to better automate tasks as a start (Makala and Bakovic, 2020).

AI product safety and liability challenges

New technological developments pose challenges to the definition of products (the extent to which it includes software, whether it is sold with the product or subsequently downloaded or updated). New technologies also bring new risks, such as a product becoming dangerous because it does not have sufficient safety protection. These new product features will complicate the task of market surveillance authorities, as well as complicating product recall, to the detriment of consumer trust.

The ambiguity in responsibility among the various economic operators in the value chain presents a problem because embedded software and other special features are an integral component of many products in circulation, which receive updates after the product has been placed on the market (European Commission, 2020b). Should the final product malfunction because of software faults, for example, there is a risk that the liability will fall on the final producer rather than on the external third-party software provider. Further complexity is added by AI and machine learning systems. The integration of an AI system that learns over time could potentially result in changes to the characteristics and functions of a product throughout its lifetime. Indeed, several EU Member States have adopted legislation on the liability of self-learning algorithms (European Parliament, 2020) (see the corresponding section). This raises the question as to how the producers of products incorporating emerging digital technology should be liable for damage caused by defective products, even where these are caused by additions that are outside of the producer’s control. It also raises the broader question of how the degree of liability should be apportioned to different types of economic operators within value chains (e.g. final producers, components and parts manufacturers, software and app developers) (Maughan, 2020).

European Commission (2020b) examines the gaps in the EU product liability and national civil liability frameworks and safety standards throughout the Union. The report also notes that the future product liability regime in the EU will increasingly have to contend with a series of key features that will affect the production and distribution of products with new embedded technologies.

[53] https://www.theverge.com/2019/2/26/18241632/google-deepmind-wind-farm-ai-machine-learning-green-energy-efficiency

Product safety and product liability provisions complement each other to ensure that risks of harm are minimised throughout the EU, and, should harm arise, victims are compensated for that damage. In terms of AI safety challenges and potential European legislation, **connectivity, opacity, data dependency** and **autonomy** should be considered in order to update the regulatory framework (Zisov and Targov (n.d.)).

Connectivity challenges the 'traditional concept of safety' as it is a direct entry for cybersecurity risks, through, for example, third party unauthorised access connected to the AI assistant. The EU Cybersecurity Act addresses these risks through a certification framework for products, services and processes. Self-driving vehicles, for instance, utilise connectivity and AI technology to navigate, and any loss of connectivity could end in possibly fatal road accidents. The General Product Safety Directive does not provide specific requirements against security threats that may affect the safety of users. The concept of safety is linked to the use of the products (legal certainty), as well as the risks that need to be addressed to make products safe for consumers under the connectivity umbrella. Connectivity brings liability challenges, for example identifying the person or organisation liable when there are several providers in a complex and connected system. Due to the Directive's technological neutrality, as well as a broad definition of 'producer', people who have suffered damage do not have a clear understanding of the person/organisation against which the consumer should direct their claim. AI systems, in particular, are set up in multiple phases (i.e. emergence, adoption, dispersion) by various actors (e.g. coders, developers, testers, users, corporations), complicating liability scrutiny.

'The **opacity** of AI systems is a major concern in terms of guaranteeing the safe exploitation of products placed within the market' (Zisov and Targov, n.d.). The opacity of AI refers to the difficulty in identifying the steps that the computer algorithms carried out independently. European product safety legislation does not address AI safety risks due to the opacity of their systems. The liability challenge is that opaque systems do not allow for the identification of the causal link between the damage and an algorithmic decision. As yet, AI algorithms are insufficiently transparent and their robustness and accountability are not regulated by a formal requirement. Human oversight is necessary as it ensures that AI software in products does not cause adverse effects. Zheng et al. (2020) underline that machine learning has made it possible 'to learn causal models from observational data' linked to decision-making. Liability law is thus affected by the opacity of AI, i.e. the underlying algorithms are no longer just codes, but rather 'black boxes' that have evolved through supervised and/or unsupervised learning processes. In conclusion, the liability challenge of opaque AI systems lies in the fact that it is difficult to pinpoint a human decision-maker (e.g. operator) for harm because machines and algorithms cannot be held liable (they do not have legal personality).

The question of whether it is possible to establish a link with human behaviour in order to constitute a liability claim when an AI system operates with a high level of autonomy is an important one. Indeed, a liability claim rests on understanding who/what was in control of the risk that materialised. AI systems can often act **autonomously (with little to no human intervention)** once they have learned the environment and the necessary tasks. Product safety rules are followed by manufacturers when considering the use of the system, providing them with legal protection. Nevertheless, safety risks may arise from the autonomous actions of the AI system, which may 'self-learn' rules that enable it to carry out an activity without human oversight. 'Autonomy can affect the safety of the product, because it may alter a product's characteristics substantially, including its safety features' (European Commission, 2020b). An AI system may necessitate an ex post risk assessment. Kamensky (2020) points to the liability imposed on manufacturers in terms of AI in medical devices and services, given the difficulty in foreseeing the AI system's behaviour and actions in a real-world medical setting. Manufacturers may thus be asked to present users with the potential unsafe actions. For a successful liability claim against producers under the Product Liability Directive, the

victim has to establish a 'defect' and the 'causal link with the damage'. For those defects that were not detectable according to the state of scientific knowledge at the time of putting the product into circulation, the Directive provides for a reverse burden of proof (i.e. the producer needs to prove that the item was safe and not defective).

The question of AI **product safety and liability** is an important reality. Beglinger (2019) has researched product liability in the field of surgical robotics and presents cases of harmful surgeries due to the malfunctioning of the technology at hand. The author argues that courts should 'infer a product defect from the occurrence of a malfunction in the absence of abnormal use and raise a rebuttable presumption that there were no reasonable secondary causes of the malfunction' (p.1044). Sullivan and Schweikart (2019) also examined the liability doctrines that address injury caused by AI and found that the current legal models are not sufficient for today's possible malpractice claims for harm caused by an AI system. Product liability in terms of software defects in AVs also poses challenges to the users to bring product liability claims against manufacturers and developers (e.g. Kim, 2018; Dempsey, 2020). Manufacturers currently do not have incentives to enhance the safety of vehicles. Arnold et al. (2019) argue that manufacturers often use so-called Supplier's Declarations of Conformity (SDoCs) to make the product worthy of consumer trust. The authors build on these SDoCs and argue that FactSheets that include explanations of the purpose, performance, safety, security and provenance information on the AI system enhance user trust. In addition to potentially proving liable and unsafe, the fact that AI services' manufacturers do not communicate the issues that cause a lack of human trust impedes broad AI adoption. Indeed, the existing liability system is not appropriate to ensure that the responsible entity that caused harmful conduct and/or harm is correctly determined in a court case (Erdélyi and Erdélyi, 2020). A victim currently has diverse liability claims, regulated at the EU level by the Product Liability Directive and the General Product Safety Directive. Nevertheless, the Directives' application in national legislation may differ slightly, meaning that claims are also regulated by the applicable national laws. These claims may be made against various persons (e.g. operator, owner, developer), which could require the victim to prove the fault of those persons, as well as a causal link (in the case of fault-based claims) or otherwise (in the case of strict liability). In the specific case of digitalisation, products, services and the value-chain are increasingly complex. Villasenor (2019) explains that the 'blame-game' goes in several directions: blaming the AI, blaming the data, blaming the users, or blaming the upstream or downstream supply chain in the case of an AI product safety liability case. Rachum-Twaig (2020) similarly identified manufacturers (designers), operators and end-users as the involved stakeholders. The large ecosystem and plurality of actors make it difficult to assess the origin of the damage. The burden of proof regarding fault and causation could be applied to these various stakeholders, as it is difficult for victims to acquire all information necessary to make a successful claim and base themselves on the correct liability framework.

Civil liability in Member States' strategies

EU Member States have adopted or are adopting strategies relating to AI. Civil liability rules are primarily regulated at Member State level. Table 5, taken from EPRS (2020) and modified by the authors, outlines the national rules on civil liability for damage caused by AI in national political initiatives. No national AI liability legislation is yet in place.

Table 4 - National rules on civil liability for damage caused by AI

Country	National strategy or policy initiative on AI	Proposals and/or national strategy provisions on liability
Belgium	Report on the impact, opportunities, possibilities and risks of the digital smart society (2019)	The report highlights the urgent need for a legislative framework, preferably at the international/EU level, and stresses that there is currently a legal deficit because there is little legislation on liability in this context.
Czechia	National AI Strategy of the Czech Republic (2018)	The strategy states that further research and analysis on liability questions are to be completed by 2021. On 10 February 2020, the Czech Office of the Government, the Czech Ministry of Industry and Trade, and the Institute of State and Law of the Czech Academy of Sciences officially launched the expert platform and forum for law and artificial intelligence (AI Observatory and Forum (AIO&F).
Estonia	AI Policy Estonia, Future of Life Institute (2019)	Legislative proposals on legal aspects of AI, including on non-contractual liability for damage caused by AI, are expected to be made in the course of 2020.
France	AI for Humanity: French Strategy for Artificial Intelligence (2018)	The national strategy is based on the 2018 Villani report. In relation to civil liability, the report proposes to (a) provide a framework for the use of predictive algorithms in such a way that a human can be held responsible at each stage of the reasoning process; (b) clarify the system of medical liability for healthcare professionals when using AI technologies; (c) define the liability regime for damage caused by the use of machine learning systems.
Germany	Artificial Intelligence Strategy (2018)	In a publication of April 2019, the Federal Ministry for Economic Affairs and Energy expressly rejected the need for additional rules on civil liability for AI. (The Bundestag has set up a committee of inquiry on AI, tasked with examining questions of liability and responsibility for AI. In the context of healthcare, the committee has recommended introducing a common certification for AI medicinal products and assessing whether there are liability gaps not covered by the general rules).

Hungary[54]	Hungarian AI Strategy 2020-2030	The strategy addresses ethics and liability by stating reliance on EU legislation.
Netherlands	Strategic Action Plan for Artificial Intelligence (2019)	The action plan stresses that it is necessary to tackle questions on liability concerning AI that present cross-border impacts at the Union level.
Slovenia	National Programme on AI	There are seven elements on which this strategy is being developed, including liability.
Sweden	National Approach for Artificial Intelligence (2018)	Sweden particularly encourages legal development in the AI area in Union law. In addition, the Swedish government is of the opinion that it would be counterproductive to adopt new laws concerning AI when the area is changing so rapidly, and instead proposes other normative tools.

Source: EPRS (2020). Civil liability regime for Artificial Intelligence.

Measures to reduce safety risks in AI applications

AI technologies bring numerous benefits, including further security. As a new technology, AI is continuously researched and implemented, with new knowledge arising through every test. This also means that multiple risks arise, either accidentally or by intentional misuse (Dafoe and Zwetsloot, 2019).

General measures in the steps of the AI value chain

McKinsey (2019) highlights the risks that arise in the lifecycle of an AI system (from conception, to use, to monitoring). In general, many risks can be mitigated by independent testing and verification throughout the operating AI system, as well as continuous reporting and analysis. The five main clusters of risk are listed below.

1. **Conceptualisation**: Conceptualisation is the starting point of understanding whether deploying an AI system in specific circumstances would be appropriate. This would be followed by setting specific rules for high-risk areas and entirely excluding certain areas in which AI systems would be allowed to be deployed (see Annex 1). To respond to potentially wrong use cases of AI systems, control examples would involve independent reviews and a set definition of an approved AI system. Feedback requirements and loops must be built into the model-development lifecycle, along with systematic tracking and reporting.
2. **Data management**: Data quality requirement and automatic anomaly detection offer the potential to avoid risks of incomplete or inaccurate data. Sensitive data should be protected, encrypted and/or masked to avoid download. This oversight measure is

[54] Added from study team research – Hungarian strategy is not reflected in the document of the European Parliament.

especially important, with poor quality data costing businesses USD 3.1 trillion per year in the US alone (Redman, 2016).

3. **Model development**: Data may be non-representative, therefore guidelines for selecting training datasets and/or algorithm testing are examples of control measures. Model outcomes may become biased or discriminatory and close attention needs to be paid to providing statistically significant input variables and independently reviewing model results to reduce bias to the greatest extent possible. A model may be unstable or degrade in performance over time – this can be avoided by periodic assessments of the models for specific degradations. Software testing usually includes unit, regression or integration tests. The performance of a machine learning model may be checked in various ways, including the Confusion Matrix, F1 Score, PR (Precision-Recall) curve, and ROC (Receiver Operating Characteristics) curve (Nighania, 2018).
4. **Model implementation**: An AI system should not be marketed too quickly, as to do so risks implementation errors. Pilots, expert testing, model testing and user testing should all be applied to avoid these errors, especially in environments that rely on connectivity (among others). Once implemented, the experts and authorities that work with the AI system may tend to follow its recommendations and allow its actions blindly, leading to serious problems. A model may be put into production differently, such as using training models (one-off, batch and real-time/online training) or serving models (Batch, Realtime (Database Trigger, Pub/Sub, web-service, inApp)) (Kervizic, 2019). The AI system developers should take into consideration the benefits and tradeoffs of the various approaches in order to avoid implementation errors.
5. **Model use and decision-making**: The codependent technological environment is necessary for the correct functioning of the AI model and infrastructure and its feeding of data. Any data that the system uses also require monitoring, at a more regular frequency for high-risk models. The correct AI model use also depends on the human-machine interface - human oversight and the possibility to change a course of action are necessary mitigation measures for the safety of AI. Any mistakes, errors, incorrect analyses and overrides should be duly tracked and reported.

Generally, an AI model can comply with international standards, such as product safety standards focusing on consumers, health software, and equipment. Standards that are under development include IEEE 7000 standards on the overview of trustworthiness in AI, and more specific standards on Big Data. One upcoming specific standard is the ISO/IEC AWI TR 5469 Artificial intelligence — Functional safety and AI systems[55].

A further relevant measure for AI is the so-called Algorithmic Impact Assessment (AIA). As underlined by the European Parliament (2019), an algorithmic system carries with it 'potentially significant consequences for individuals, organisations and societies' given their use in both the public and private sectors. An AIA is a requirement for any system that may cause a 'potentially severe non-reversible impact', be it in the private or public sector. A regulatory body for algorithmic systems should be tasked with auditing the AIAs of systems requiring high-level oversight, such as those used in highly sensitive and/or safety-critical application domains (e.g. healthcare).

[55] For more on the standard, see https://www.iso.org/standard/81283.html?browse=tc

Safety measures for an AI system

The self-learning and autonomous behaviour of AI systems entails potential impacts on the safety of a product, requiring a new risk assessment. From the outset, human oversight as a principle is a key safeguard to mitigate safety risks in AI systems.

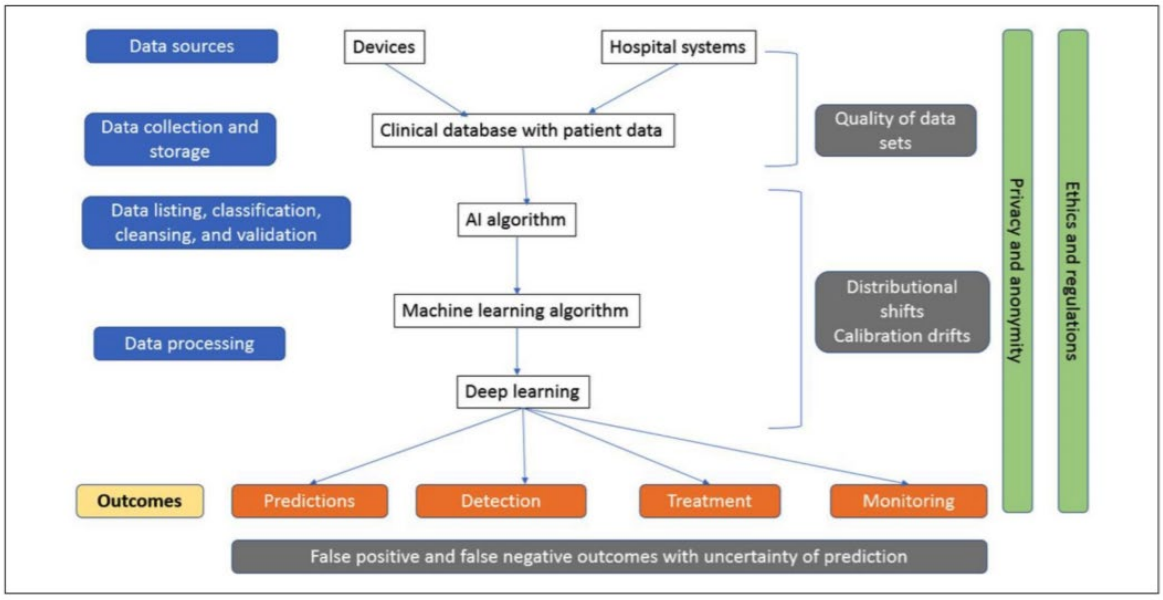
The AI HLEG (2020) published an Assessment List for Trustworthy Artificial Intelligence (ALTAI). One key element to mitigate safety risks is to ensure that appropriate human oversight measures are in place through governance mechanisms. The ALTAI lists the following criteria for AI system developers and deployers to assess whether or not human oversight is in place:

- 'Have the humans (HITL, HOTL, HIC) been given specific training on how to exercise oversight?
- Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?
- Did you ensure a "stop button" or procedure to safely abort an operation when needed?
- Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?'

Human oversight may involve the following further elements:

1. Ensuring that the data source is trustworthy (this is the most important step, as there may underlying bias in the data itself, for example).
2. Ensuring that a human can lead the decision and not the machine (which is the largest source of safety risks).
3. Carrying out a conformity assessment to check the global impact of an AI product (i.e. determining how safe is 'safe'? Are there fewer safety risks with the AI system than by human error?).

The following section focuses on safety issues in the healthcare context, drawing up the value-chain of an AI project to identify where safety issues could arise. Ellahham et al. (2019) illustrate the safety concerns at various stages of deployment of AI in healthcare.



The high quality of datasets is key in training AI systems in healthcare. Any failure in the initial data may cause incorrect outcomes and function erroneously throughout its application period, invalidating the entire AI system. Ensuring that the data source is trustworthy and correct is key to preventing safety issues caused by AI. Human bias within the training dataset is a common issue in automated systems and it can compromise the safety of an AI system.

Unexpected behaviours and unscalable oversight are to be expected in any AI system. The self-learning phase of AI is difficult to predict, especially if it is not done on a large-scale and in a 'perfectly' predictable environment during the supervised learning algorithm inputs. Indeed, training samples and test samples may differ, especially once the system is put into the 'real world'. As developers are required to monitor how algorithms are changing, a reassessment of the data is necessary if there are any unexpected and/or unwanted changes (along with the notification of the necessary bodies) (Reardon, 2019). Unsupervised machine learning algorithms and systems are prone to attacks that may impact their safety, such as adversarial attacks, data poisoning and model stealing.

Elahham et al. (2019) describe four strategies to ensure the safety of AI systems in healthcare:

- Safe design: testing and ensuring that there are no potential safety hazards in the AI system.
- Safety reserves: detection of uncertainty in training.
- Safe fail: ensuring a back-up mode in the system in case the first intended deployment fails.
- Procedural safeguards: including a user-experience design.

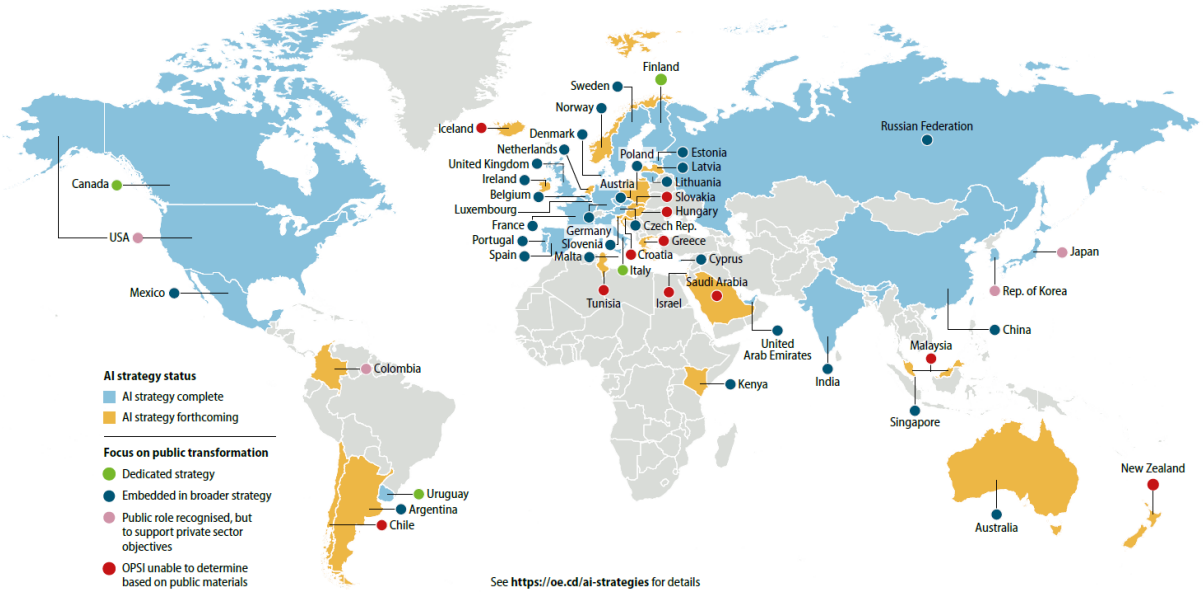
In healthcare (and other sectors), human oversight must control the AI system. In medicine, software may be used for various purposes, including the dosage of medicines that need to be adapted to the patient or to time, which is a difficult learning curve for AI. An updating protocol for any new software calibrations requires a new assessment, while the costs, risks and uncertainties should be defined for every application.

Safety data disclosure is necessary, as are privacy and sharing of data related to the use of AI applications in healthcare, and high normative standards defined and adopted throughout the lifecycle of the AI product. Health systems are complex and involve a wide range of actors and institutions. The development and deployment of AI systems should be assembled in collaboration with data scientists and clinical staff in order to ensure proper knowledge transfer. If the two parties do not interact in the initial development phase of the AI (sharing real-life experience), there may be potential missing key information in understanding hospital systems and devices, and thus how data processing is carried out in theory and in practice. 'AI actors should also be accountable for the proper functioning of their algorithms, within the scope of their own roles' (OECD, 2020). Developers are required to monitor how algorithms are changing and notify the necessary bodies of unexpected and/or unwanted changes (Reardon, 2019). Although the specific risks that providers and users should consider varying by use case, oversight of the data sources during the development phase is crucial, as is ensuring human oversight during the deployment phase.

INTERNATIONAL EXPERIENCE OF AI POLICY: EMERGING REGULATORY FRAMEWORKS

The landscape of national AI strategies is extremely rich, with several repositories of information on existing initiatives, including those of the OECD, the AI Watch of the JRC, the Future of Life Institute, AlgorithmWatch and others (see Figure 10 below). At the same time, the forthcoming EU legal act appears to be innovative in proposing a comprehensive regulatory framework for AI. Governments in third countries appear to be lagging in the development of a framework and will look to the EU as a standard setter (e.g. India, Japan), will be less eager to take action to impose regulatory constraints on AI (e.g. China), or will be more inclined towards sectoral approaches, rather than all-encompassing frameworks (US). A lively debate is envisaged on the possibility of legislation in the domain of AI in many countries. Specific national experiences are outlined below, in which some of the regulatory requirements mentioned in the White Paper and in the Inception Impact Assessment are considered.

Figure 9 - International landscape of AI initiatives

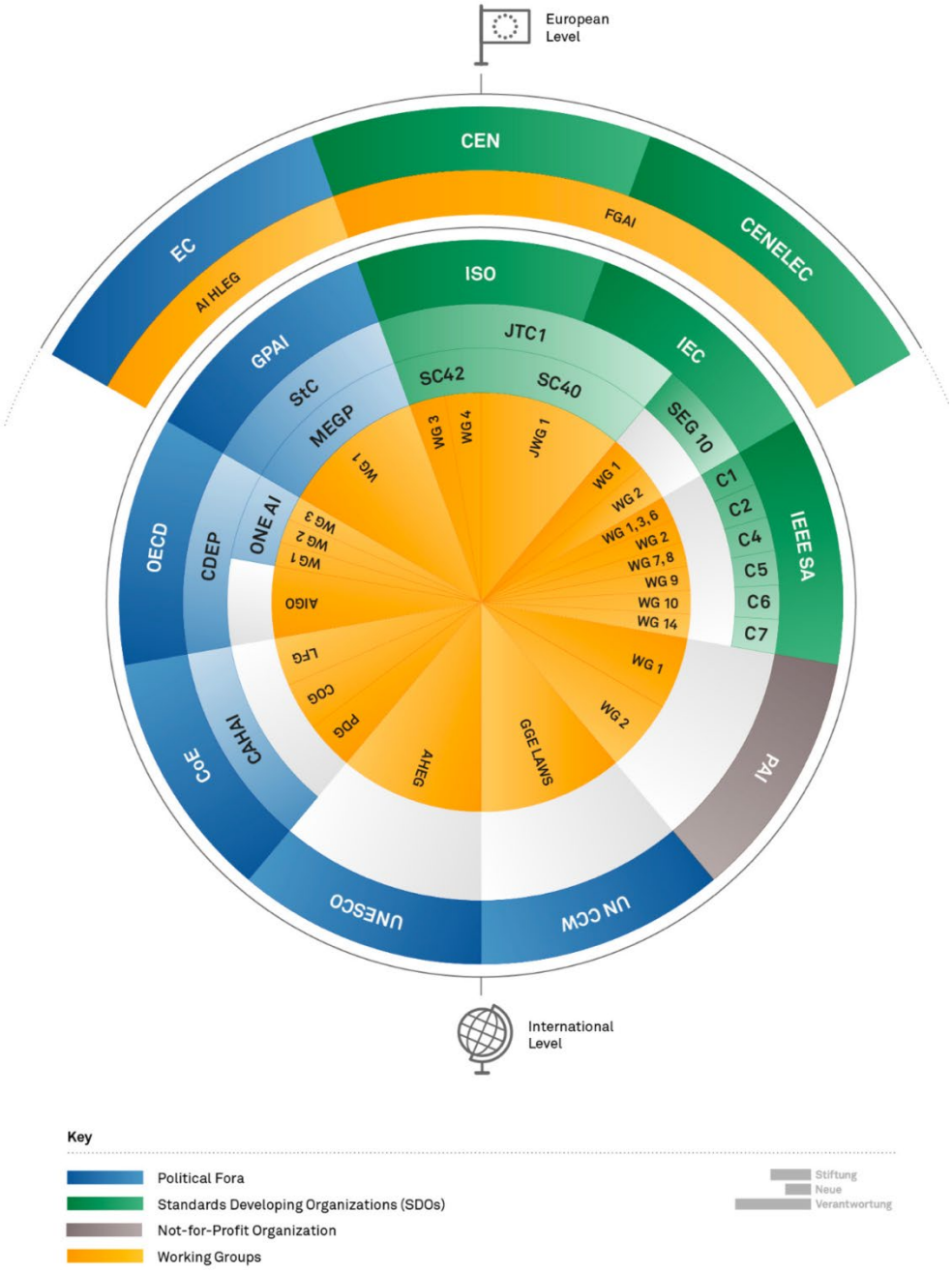


Source: OECD OPSI

1. Emerging policy approaches to AI risks: scope, requirements and governance

To date, no country has enacted a comprehensive regulatory framework on AI. In some countries, however, the debate has advanced such that a definition of AI, a governance framework for overseeing its development and diffusion, and some requirements for its trustworthiness have been either proposed or implemented. The initiatives adopted in Australia, Canada, Japan, Singapore, the UK and the US are described below.

Figure 10 - Stakeholders engaged in AI governance



Source: Stiftung Neue Verantwortung (2020)

a. Australia's voluntary framework

The Australian government is developing a voluntary AI Ethics framework, with the following characteristics.

Definition of AI. AI is defined as 'A collection of interrelated technologies used to solve problems autonomously and perform tasks to achieve defined objectives without explicit guidance from a human being' (Dawson et al. 2019). This definition is thus technology-neutral and very broad, and covers both recent, powerful advances in AI (such as neural nets and

deep learning), as well as less sophisticated applications with significant impacts on people, such as Automated Decision Making systems (ibid.).

- Eight voluntary AI Ethics principles as stated by the Australian Ombudsman are (Ombudsman Australia n.d.):
    - Human, social and environmental wellbeing: Throughout their lifecycle, AI systems should benefit individuals, society and the environment.
    - Human-centred values: Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals.
    - Fairness: Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.
    - Privacy protection and security*:* Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.
    - Reliability and safety*:* Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose.
    - Transparency and explainability*:* There should be transparency and responsible disclosure to ensure that people know when they are being significantly impacted by an AI system, and can find out when an AI system is engaging with them.
    - Contestability: When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or output of the AI system.
    - Accountability: Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

Guidance has been developed to help businesses to apply these principles in their organisations. The Australian government clarifies that not every principle will be relevant to every use of AI. For instance, many businesses use systems that may include AI. Examples are email or accounting software and the use of AI in these systems is unlikely to be sufficiently impactful to require the use of the principles. Importantly, if a specific AI use does not involve or affect human beings, organisations may not need to consider all of the principles (Australian Government 2019).

- Addresses. The framework is addressed to both developers and implementers (e.g. deployers) of AI systems. The guidance prompts them to consider two main questions when developing or implementing AI:
    - “Will the AI system you are developing or implementing be used to make decisions or in other ways have a significant impact (positive or negative) on people (including marginalised groups), the environment or society?”
    - “Are you unsure about how the AI system may impact your organisation or your customers/clients?”

A 2019 public consultation sparked off a lively discussion, with 130 written submissions related to the voluntary framework (among others) (Australian Government 2019). The Australian

government summarised the results of the consultation in a dedicated [56]. The government reported several important results: considerable support for a principles-based framework that can guide the design, development, deployment and operation of AI in Australia; the need for an iterative and flexible framework to ensure that it adapts to technological change; the lack of reference to 'security' and the need for more consideration of diversity and human oversight; and the need for a risk-based regulatory framework based on careful consideration of existing regulatory gaps.

On individual requirements the Australian government writes:

- Interesting feedback was received on the principle of *fairness*, with stakeholders advocating a sharper definition and more focus on avoiding discrimination of minority groups, the inclusion of concepts of inclusion and accessibility, and a broader focus - fairness should not only be limited to algorithms and training data, and needs to be considered over the entire lifecycle of an AI system.
- On *transparency*, stakeholders found that the principle may be challenging to apply in practice, due to the difficulty of explaining AI systems and decisions in an easily understandable way. It should also ensure that people are provided with a reasonable justification of the outcome from the AI system in a user-friendly format, and that requirements for explainability are applied in a way that is proportional to the potential risks and impact of a given AI system.
- The principle of *contestability* saw stakeholders advocate for more guidance and the need to clearly communicate that redress is possible when things go wrong, as a vital aspect of building public trust in AI.
- On *accountability*, comments addressed the need for improved clarity on who would be considered accountable, especially in the case of open-source algorithms, and when AI systems are used beyond their original intent. Accountability should focus on the outcomes of AI systems and ensure appropriate levels of human oversight.

The release of the AI Ethics framework and the release of the AI Roadmap by Australia's data innovation network (Data61, hosted by the national science agency) was accompanied by initiatives at the subnational government level, in particular the New South Wales Government's AI Ethics Framework[57]. The Australian Human Rights Commission continues to enquire into the human rights impacts of new technologies, specifically AI. The recently published report by ACOLA (Australian Council of Learned Academies) provides an in-depth horizon scan of AI, including areas like issues of algorithmic fairness more generally, or the intersection between AI and the rights of indigenous peoples (ACOLA, 2019).

Finally, the development of standards related to AI has been subject to important institutional reflection in Australia, with Standards Australia (the national standards body) holding national consultation forums and deep-dive workshops across major capital cities in 2019, concluding with an AI Standards Lab to test key ideas and the subsequent presentation of a Standards Roadmap in February 2020. Those consulted reportedly underlined the 'opportunity that exists to turn salient concerns into opportunities to develop "responsible AI" by tackling specific concerns in areas such as privacy, inclusion, safety and security and getting the policy and regulatory balance right' (ibid.). Realising this opportunity will require effective national co-

[56] For consultation summary, see: https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/developing-the-ai-framework-and-principles

[57] For New South Wales AI Ethics framework, see: https://www.digital.nsw.gov.au/transformation/policy-lab/artificial-intelligence-ai/nsw-ai-ethics-framework

ordination, a task for both Australian businesses and government, with the support of Standards Australia.

b. Canada's Directive on Automated Decision-Making

Canada is one of the most active countries in the definition of principles for responsible AI development and has adopted a sectoral approach to the definition of a regulatory framework for AI. The Personal Information Protection and Electronic Documents Act (PIPEDA) is a federal privacy law addressed in the private sector. The House of Commons Standing Committee on Access to Information, Privacy and Ethics released a report on 28 February 2018 in which they recommend adapting the legislation in a manner that is 'heavily influenced by the direction set in the European Union General Data Protection Regulation' (Office of the Privacy Commissioner of Canada, 2017). The report expressed concerns about the transparency of AI decision-making and the risk of algorithms using personal information to 'perpetuate prejudices or discriminatory practices', and recommended that 'the Government of Canada consider implementing measures to improve algorithmic transparency.' (House of Commons Standing Committee on Access to Information, Privacy and Ethics, 2018)

Under the authority of the Financial Administration Act, the Treasury Board of Canada issued a Directive on Automated Decision-Making, which took effect on 26 November 2018. The Directive lists the responsibilities of federal institutions deploying AI-automated decision systems and aims to facilitate the ethical and responsible use of AI. The Directive came into effect on 1 April 2020 and applies to the use of ADM systems that 'provide external services and recommendations about a particular client, or whether an application should be approved or denied' (Zhu, 2018). It includes an AIA designed 'to help [federal institutions] better understand and reduce the risks associated with Automated Decision System's,' which should be carried out 'prior to the production of any Automated Decision System' by the federal authorities (Government of Canada, 2018). Pursuant to the Canadian Policy on the Management of Information Technology, 'this Directive does not apply to any national security systems' (ibid.).

Some of the Directive's main characteristics are outlined below. The following text provides excerpts from the Directive on Automated Decision-Making (Government of Canada, 2018).

- *Definition of AI.* AI is defined as 'Information technology that performs tasks that would ordinarily require biological brainpower to accomplish, such as making sense of spoken language, learning behaviours, or solving problems.' (ibid.) The definition is therefore technology-neutral, and oriented towards encompassing all ADM. The latter is defined as including 'any technology that either assists or replaces the judgement of human decision-makers. These systems draw from fields like statistics, linguistics, and computer science, and use techniques such as rule-based systems, regression, predictive analytics, machine learning, deep learning, and neural nets' (ibid.).
- *(Selected) Key provisions.* The Assistant Deputy Minister responsible for the programme using the ADM, or any other person named by the Deputy Head, is responsible for carrying out an AIA prior to the production of any ADM system. The AIA applies a number of requirements, depending on the risk classification of the ADM system, resulting in different 'levels' of AIA (see below). The AIA must be updated 'when system functionality or the scope of the Automated Decision System changes' (ibid.), although no further guidance is provided as to what constitutes such change. The final results of AIAs must be released in an accessible format. Importantly, the Directive requires institutions to '[provide] notice on relevant websites that the decision rendered will be undertaken in whole or in part by an Automated Decision System'' (ibid.), and that a meaningful explanation is provided to affected individuals on how

and why the decision was made. 'The Government of Canada retains the right to access and test the ADM system, including all released versions of proprietary software components, in case it is necessary for a specific audit, investigation, inspection, examination, enforcement action, or judicial proceeding, subject to safeguards against unauthorised disclosure.' (ibid.) It also 'retains the right to authorise external parties to review and audit those components as necessary' (ibid.). Before production, processes must be developed so that the 'data and information used by the ADM system are tested for unintended data biases and other factors that may unfairly impact the outcomes' (ibid.). The data collected must be validated to ensure that it is 'relevant, accurate, up-to-date, and in accordance with the Policy on Information Management and the Privacy Act.' (ibid) Deputy ministers are also responsible for 'conducting risk assessments during the development cycle of the system and [establishing] appropriate safeguards to be applied, as per the Policy on Government Security.' (ibid.) They should also ensure that ADM systems used in government allow for human intervention, when appropriate.

- *Risk classification*. The Canadian Directive contains several levels of AIA, as outlined in Table 6 below.

Table 5 - Risk classification in the Government of Canada Directive on Automated Decision-Making

Level	Description
I	The decision will likely have little or no impact on: • the rights of individuals or communities, • the health or wellbeing of individuals or communities, • the economic interests of individuals, entities, or communities, • the ongoing sustainability of an ecosystem. Level I decisions will often lead to impacts that are reversible and brief.
II	The decision will likely have moderate impacts on: • the rights of individuals or communities, • the health or wellbeing of individuals or communities, • the economic interests of individuals, entities, or communities, • the ongoing sustainability of an ecosystem. Level II decisions will often lead to impacts that are likely reversible and short-term.
III	The decision will likely have high impacts on: • the rights of individuals or communities, • the health or wellbeing of individuals or communities, • the economic interests of individuals, entities, or communities, • the ongoing sustainability of an ecosystem. Level III decisions will often lead to impacts that can be difficult to reverse and are ongoing.

IV	The decision will likely have very high impacts on: ● the rights of individuals or communities, ● the health or wellbeing of individuals or communities, ● the economic interests of individuals, entities, or communities, ● the ongoing sustainability of an ecosystem. Level IV decisions will often lead to impacts that are irreversible and are perpetual.

Source: Government of Canada Directive on Automated Decision-Making, Appendix B

This risk classification results in different impact level requirements, as shown in Table 7 below.

Table 6 - impact level requirements in the Canada Directive on Automated Decision-Making

Requirement	Level I	Level II	Level III	Level IV
Peer review	None	**At least one of:** Qualified expert from a federal, provincial, territorial or municipal government institution Qualified members of faculty of a post-secondary institution Qualified researchers from a relevant NGO Contracted third-party vendor with a related specialisation Publishing specifications of the ADM system in a peer-reviewed journal A data and automation advisory board specified by Treasury Board Secretariat		**At least two of:** Qualified experts from the National Research Council of Canada, Statistics Canada, or the Communications Security Establishment Qualified members of faculty of a post-secondary institution Qualified researchers from a relevant NGO Contracted third-party vendor with a related specialisation A data and automation advisory board specified by Treasury Board Secretariat **OR:** Publishing specifications of the ADM system in a peer-reviewed journal
Notice	None	Plain language notice posted on the programme or service website.	Publish documentation on relevant websites about the ADM system, in plain language, describing: ● How the components work ● How it supports the administrative decision ● Results of any reviews or audits ● A description of the training data, or a link to the anonymised training data if these data are publicly available	
HITL for decisions	Decisions may be rendered without direct human involvement		Decisions cannot be made without having specific human intervention points during the decision-making process and the final decision must be made by a human	

Explanation requirement	In addition to any applicable legislative requirement, ensuring that a meaningful explanation is provided for common decision results. This can include providing the explanation via a Frequently Asked Questions (FAQ) section on a website	In addition to any applicable legislative requirement, ensuring that a meaningful explanation is provided on request for any decision that resulted in the denial of a benefit, a service, or other regulatory action	In addition to any applicable legislative requirement, ensuring that a meaningful explanation is provided with any decision that resulted in the denial of a benefit, a service, or other regulatory action	
Testing	Before going into production, develop the appropriate processes to ensure that training data are tested for unintended data biases and other factors that may unfairly impact outcomes Ensure that data used by the ADM system are routinely tested to ensure that they are relevant, accurate and up-to-date			
Monitoring	Monitor the outcomes of ADM systems on an ongoing basis to safeguard against unintentional outcomes and to ensure compliance with institutional and program legislation, as well as this Directive			
Training	None	Documentation on the design and functionality of the system	Documentation on the design and functionality of the system. Training courses must be completed	Documentation on the design and functionality of the system Recurring training courses A means to verify that training has been completed
Contingency planning	None		Ensure that contingency plans and/or backup systems are available should the ADM system be unavailable	
Approval for the system to operate	None	None	Deputy Head	Treasury Board

Source: Government of Canada Directive on Automated Decision-Making, Appendix C

c. The German Data Ethics Commission's proposed risk classification

In September 2018, the German Federal Government set up the German Data Ethics Commission (GDEC) and tasked it with building guidelines for the safe and ethical development and use of AI systems. In October 2019, the GDEC published a report that included 75 recommendations for regulating algorithmic systems, including AI and other data technologies. The report proposes an EU Regulation on Algorithmic Systems (EU-ASR). While endorsing the EU's Ethics Guidelines for Trustworthy AI, the GDEC asserts that binding rules with concrete regulatory requirements are necessary. The proposed policy framework approach is similar to that of the GDPR, in that it focuses on individual rights and corporate accountability, and has a horizontal application.

The report puts forward a universally applicable five-level scale of 'criticality' on which different AI systems are classified according to the degree of the potential harm they create. The rationale of the approach is that the greater the risk for potential harm, the more intrusive regulatory intervention is necessary. The 'criticality system' is differentiated from the risk classification proposed by the Commission in the White Paper on AI because it departs from the high-low risk dichotomy. In addition, while one of the Commission's proposed criteria for classifying AI technologies as 'high risk' for regulation is the 'sensitivity of the sector' in which the AI application is deployed, the GDEC proposes a sector-neutral framework.

The design of the proposed regulatory framework is ex ante, periodic and ex post, and builds on both self-regulation and enforcement by supervisory authorities. The mandatory regulatory requirements with which developers and deployers of AI systems would need to comply include risk impact assessment, licensing procedures, mandatory labelling, mandatory access to information, minimum quality standards, anti-discrimination obligations, and other transparency and accountability mechanisms.

The five levels of criticality proposed by the GDEC are:

Level 1. AI systems falling under Level 1 are considered to pose zero or negligible potential for harm, thus there is no need to subject them to new regulatory requirements.

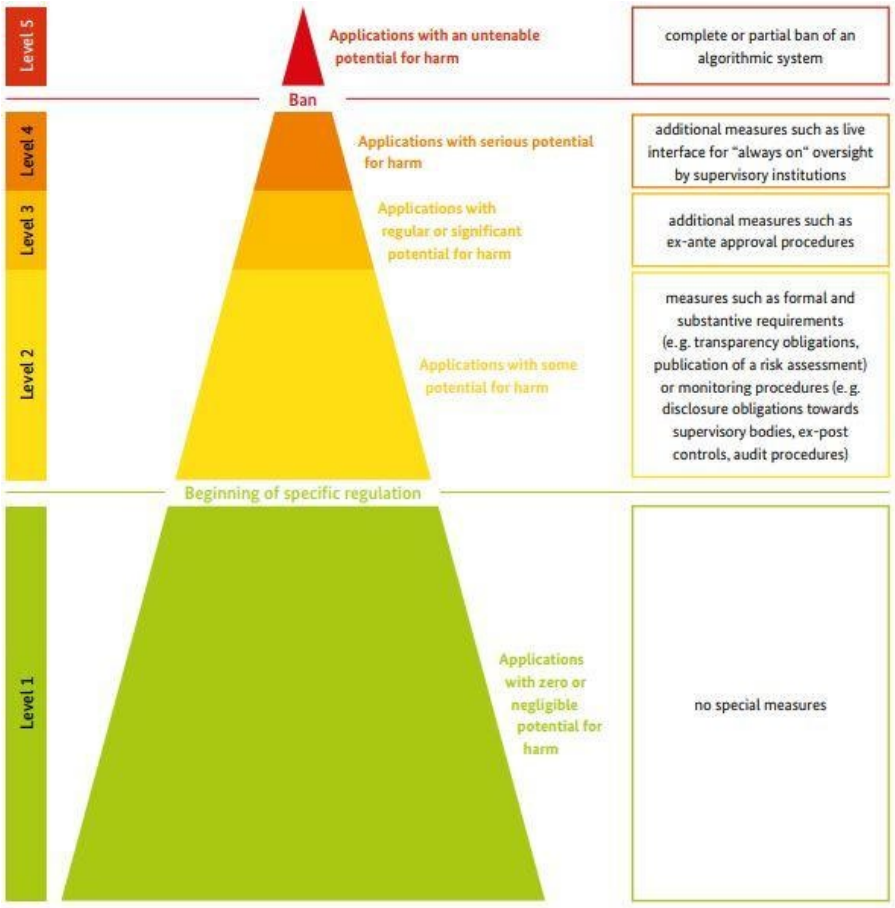
Level 2. AI systems with some potential for harm should be subject to ex post regulation, such as mandatory labelling obligations (e.g. publication of risk assessment), monitoring or reporting mechanisms (disclosure to supervisory bodies, auditing), and/or transparency requirements (e.g. right to access information).

Level 3. AI systems with regular or significant potential for harm would be subject to the requirements applying to Level 2 AI systems and additional ex ante measures, such as an approval procedure before being placed on the market.

Level 4. AI systems with serious potential for harm would be subject to increased transparency requirements and continuous market surveillance.

Level 5. AI systems falling under Level 5 are considered to pose untenable potential for ham and the GDEC thus recommends a full or partial ban.

Figure 11 - Risk levels proposed by the German Data Ethics Commission



Source: GDEC (2019)

The report also acknowledges the gaps in the current national and EU civil liability frameworks for damage caused by autonomous technology applications. The GDEC recommends revising the current system to oblige human operators of AI applications to bear vicarious (strict) liability for any potential damage.

d. Japan's Contract Guidelines on Utilisation of AI and Data

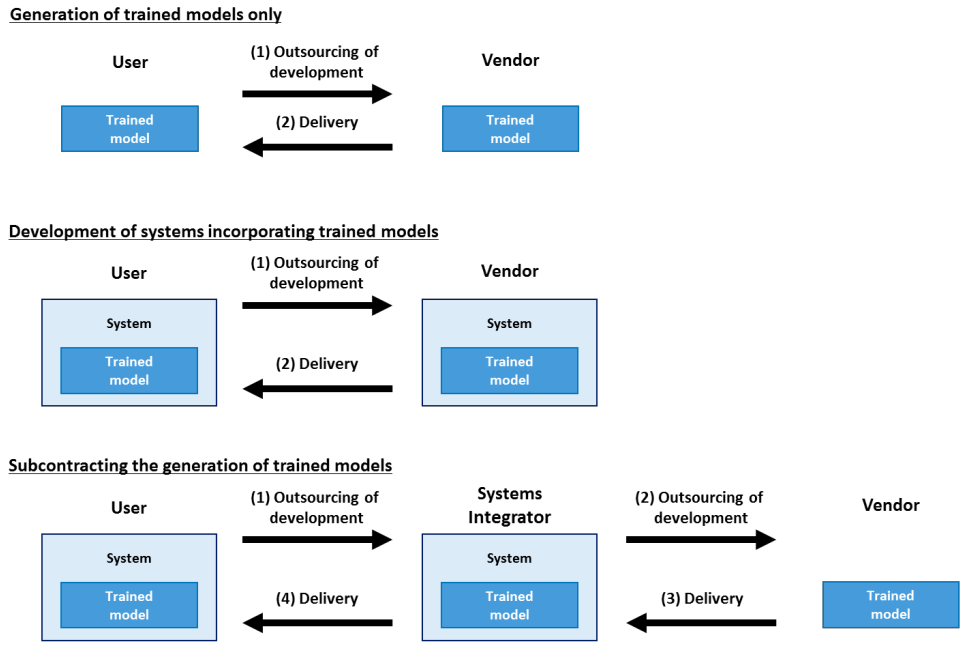
Japan is very advanced in its AI development and use, with the legislation in place to facilitate the flow and protection of data (Basic Act on the Advancement of Utilising Public and Private Sector Data; Act on the Protection of Personal Information). Japan has not taken direct action to define a regulatory framework for AI in a comparable way to the forthcoming AI legal act in the EU. However, in June 2018, the Ministry of Economy, Trade and Industry (METI) published the Contract Guidelines on Utilisation of AI and Data, 'a reference for businesses that explains approaches to concluding (i) contracts for utilisation of data, or (ii) contracts for the development and utilisation of software using AI technology.' (Japanese Ministry of Economy, Trade and Industry, 2018). The Guidelines describe the main challenges, unresolved issues, model contract clauses, elements to be considered in the preparation of contract clauses, and other key points.

The main elements that are of interest here are summarised below. The following text provides excerpts from Japan's Contract Guidelines on Utilisation of AI and Data.

- *Definition of AI*. While the Guidelines acknowledge the lack of a generally established definition of AI, they offer a 'rough' classification into '(i) general-purpose AI, based on the concept of creating machines that possess human intelligence itself ("Strong AI"), and (ii) AI based on the concept of causing machines to perform activities that humans use their intelligence to perform ("Weak AI"). The Guidelines state that for the purpose of the relevant chapters, they refer to "Weak AI" as "AI" as these technologies have 'reached the state of practical application' (ibid.) 'AI technology' is defined in the Guidelines as 'a generic term for a series of software technologies that enable computers to perform intellectual activities that can be performed by humans.' (ibid.) For convenience's sake, the Guidelines assume the term 'AI technology' to mean either 'machine learning' or a 'series of software technologies related to machine learning' (ibid.). The terms 'machine learning', 'supervised learning', 'unsupervised learning' and 'deep learning' are also defined (METI, 2019).
- *Explanation of 'how things can go wrong' in AI-related contract*s. The Guidelines are extremely detailed in the description of the different interests of parties along the value chain of AI development and deployment. They explain that the 'positions and attitudes between parties with respect to the development or utilisation of AI technology differ, so various problems can arise when a contract is executed, including: (i) problems unique to raw data (whether raw data exists, propriety of or delays in provision, quality and sufficiency, and the like); (ii) problems unique to AI-based software (whether completion is possible and whether an obligation to complete the software exists, the quality of the developed software, and the like); (iii) problems regarding ownership of intellectual property rights and terms of use (deliverables, intellectual property produced in the course of development, and AI products (outputs)); (iv) problems regarding liability; and other problems caused by inconsistencies and the like between the purpose of development and utilization (business needs) on the user side and technical knowledge on the Vendor side.' (ibid.) All of these aspects are described in detail and accompanied by model contract clauses.
- *Different models for developing AI-based contracts*. The Japanese Guidelines distinguish between three categories of use cases:
    - Categories of contracts involving generation of trained models only, which include two cases: 'when a User provides data and a Vendor individually generates a trained model only' (e.g. insurance company Y requests a data analysis company X to analyse the data of Company Y; X performs machine learning on the data and delivers to Y a trained model that possesses the requested functionalities), or cases involving the 'development of systems incorporating trained models' (e.g. equipment manufacturer X considers installation of a trained model in monitoring equipment that is provided to Company Y in order to enable detection of a specific object; the trained model is generated through training using combined image data provided by both X and Y) (ibid.).
    - Categories of contracts 'involving development of systems incorporating trained models', such as 'when a User provides data and a Vendor individually develops a system incorporating a trained model' (trading company Y provides a training dataset, and a machine learning developer X that accepts delegation from Y develops a system incorporating a trained model by using that training dataset and delivers that system to Y), or 'when a Vendor prepares data by itself and individually generates a trained model, and another business operator develops an entire system based on the trained model' (ibid.).
    - 'Categories involving subcontracting the generation of trained models', e.g. 'when a systems integrator […] that has accepted from a User the delegation

of the development of an entire system subcontracts to a Vendor the generation of a trained model only' (e.g. Y outsources the development of an identification system to a vendor X1 and a systems developer X2. X1 generates a trained model using data prepared by itself, and X2 incorporates that trained model into an identification system and delivers that system to Company Y) (ibid.).

Figure 12 - Developmental categories in Japan's Contract Guidelines on Utilisation of AI and Data



Source: METI (2019)

e. Singapore's model governance framework on AI

In January 2019, the Personal Data Protection Commission (PDPC) of Singapore published the first edition of a model AI governance framework for consultation. The framework offers detailed and practical guidance to businesses to address key ethical and governance issues when putting AI solutions into production (OECD 2020). The model contains explanations of how AI systems work, how to build good data accountability practices, and create open and transparent communication. The framework was revised on 21 January 2020. It is accompanied by an Implementation and Self-Assessment Guide for Organisations (ISAGO), developed together with the World Economic Forum's Centre for the Fourth Industrial Revolution, and in close consultation with industry, with contributions from over 60 organisations. ISAGO was further complemented by a Compendium of Use Cases illustrating how local and international organisations across different sectors and sizes have adapted their AI governance practices with all parts of the model framework. It also shows how organisations have started accountable AI governance practices, and profited from the use of AI (PDPC 2020).

The model framework is presented as algorithm-neutral (it does not focus on specific AI or data analytics methodology, and applies to the design, application and use of AI in general); technology-neutral (by reason of not targeting certain specific systems, software or technology and being agnostic towards the different types of development languages and data storage methods); sector-neutral (it serves as a checklist of considerations and measures for organisations regardless of which sector they operate in, , and invites specific sectors or

organisations to incorporate additional sector-specific considerations and measures or adjust the baseline considerations according to their needs); and scale and business-model-neutral (it does not address organisations of a specific scale or size; it can be used in B2B, B2C and other settings). The model framework is based on two high-level principles: that organisations using AI in decision-making should ensure that the decision-making process is explainable, transparent and fair, and that AI solutions should be human-centric (IMDA, PDPC 2020).

In the model framework, AI is defined as 'a set of technologies that seek to simulate human traits such as knowledge, reasoning, problem solving, perception, learning and planning, and, depending on the AI model, produce an output or decision (such as a prediction, recommendation, and/or classification). AI technologies rely on AI algorithms to generate models. The most appropriate model(s) is/are selected and deployed in a production system.' (PDPC, 2020)

Key areas covered by the model framework

The following text provides excerpts from the second edition of the Model AI Governance Framework of Singapore.

Internal governance structures and measures: Setting up (or adapting) internal governance structures and measures to incorporate values, risks and responsibilities relating to algorithmic decision-making. The key here is to allocate clear responsibilities in the organisation for the ethical development of AI[58], in particular defining arrangements for risk management and internal controls.

Determining the level of human involvement in AI-augmented decision-making: This contains a methodological framework aimed at helping organisations in 'setting their risk appetite for use of AI' , i.e. deciding what level of risks they deem appropriate and determining the degree of human oversight of the AI-augmented decision-making process (PDPC, 2020). A key message in the framework is that 'identifying commercial objectives, risks and determining the appropriate level of human involvement in AI-augmented decision-making is an iterative and ongoing process' (ibid.). Accordingly, the model framework draws attention to the need to periodically and continuously review the risks associated with different technological solutions, implement mitigating measures and provide an appropriate response plan should the risk materialise 'Documenting this process through a periodically reviewed risk impact assessment helps organisations to develop clarity and confidence in using the AI solutions, and aids them in responding to potential challenges from individuals, other organisations or businesses, and regulators.' (ibid.)

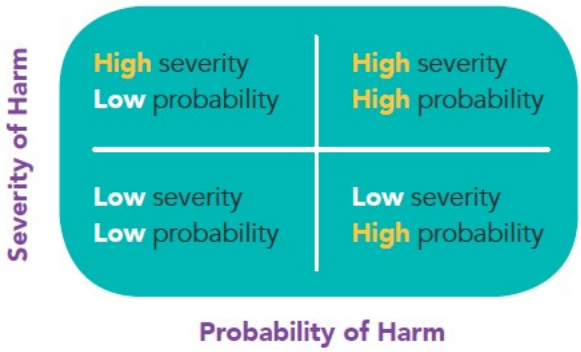
Like the AI HLEG Ethics Guidelines for Trustworthy AI, the model framework distinguishes between different arrangements for human oversight, including HITL ('human oversight is active and involved, with the human retaining full control and the AI only providing recommendations or input; decisions cannot be exercised without affirmative action by the human, such as a human command to proceed with a given decision'); HOTL ('there is no human oversight over the execution of decisions; the AI system has full control without the

[58] For example, using any existing risk management framework and applying risk control measures to assess and manage the risks of deploying AI, including any potential adverse impact on the individuals, decide on the appropriate level of human involvement in AI-augmented decision-making and manage the AI model training and selection process; maintain, monitor, document and review the AI models that have been deployed, with a view to taking remediation measures where needed; review communications channels and interactions with stakeholders to provide disclosure and effective feedback channels; and ensure that relevant staff dealing with AI systems are properly trained.

option of human override'); and human-over-the-loop ('human oversight is involved to the extent that the human is in a monitoring or supervisory role, with the ability to take over control when the AI model encounters unexpected or undesirable events such as model failure') (ibid.).

The model framework uses a matrix to guide organisations in the selection of the best governance model, which bases the risk on the probability and severity of harm (Figure 14). Such an approach was also proposed in several contributions to the consultation on the White Paper. The PDPC explains that the probability and 'severity of harm are [not] the only factors to be considered in determining the level of human oversight in an organisation's decision-making process involving AI'. Other factors could include 'the nature of harm (i.e. whether the harm is physical or intangible in nature)', 'the reversibility of harm' (and as a corollary, the ability for individuals to obtain recourse), and 'whether it is operationally feasible or meaningful for a human to be involved in a decision-making process (e.g. having a human-in-the-loop would be unfeasible in high-speed financial trading, and be impractical in the case of driverless vehicles' (ibid.). In the model, organisations working on AI solutions for safety-critical systems are encouraged to 'ensure that a person be allowed to assume control, with the AI system providing sufficient information for that person to make meaningful decisions or to safely shut down the system where human control is not possible' (ibid.).

Figure 13 - Severity and probability of harm in Singapore's model AI governance framework



Operations management: The Model Framework lists the 'Issues to be considered when developing, selecting and maintaining AI models, including data management' (ibid). The latter is detailed through a number of good data accountability practices, which include understanding the lineage of data ('where the data originally came from, how it was collected, curated and moved within the organisation, and how its accuracy is maintained over time' (using approaches such as backward, forward or end-to-end data lineage)), ensuring data quality (accuracy, completeness, veracity, relevance, integrity, usability of the dataset, and human interventions (e.g. if any human has filtered, applied labels, or edited the data)), minimising inherent bias (in particular selection bias (e.g. omission bias, stereotype bias)) and measurement bias, using different datasets for training, testing, and validation, and periodically reviewing and updating the datasets. This section of the model framework also details 'numerous features or functionalities enabled through algorithms in AI models', including measures such as 'explainability, repeatability, robustness, regular tuning, reproducibility, traceability, and auditability, which can enhance the transparency of algorithms found in AI models'[59] (ibid.). The framework encourages organisations to adopt a risk-based

[59] The framework clarifies that 'It may not be feasible or cost-effective to implement even the most essential of these measures for all algorithms.' It also adds that 'some of these measures like explainability (or repeatability, when using models that are not easily explained), robustness and regular tuning are sufficiently essential that they

approach by carrying out a two-step assessment: (i) 'identify the subset of features or functionalities that have the greatest impact on stakeholders for which such measures are relevant'; and (ii) 'identify which of these measures will be most effective in building trust with stakeholders' (ibid.).

Stakeholder interaction and communication: The Model Framework describes communication strategies to be used with the organisation's stakeholders and for the management of relationships with those.. The model framework encourages organisations to 'provide general information on whether AI is used in their products and/or services', including (where appropriate) 'information on what AI is, how AI is used in decision-making in relation to consumers, what are its benefits, why your organisation has decided to use AI, how your organisation has taken steps to mitigate risks, and the role and extent that AI plays in the decision-making process' (ibid.). Organisations are also invited to 'consider disclosing the manner in which an AI decision may affect an individual consumer, and whether the decision is reversible' (ibid.). Interestingly, the model framework invites organisations to consider providing consumers with an option to opt-out of AI-enabled decisions, depending on a number of factors (degree of risk/harm to the individuals; the reversibility of the decision made; the availability of alternative decision-making mechanisms; the cost or trade-offs of alternative mechanisms; the complexity and inefficiency of maintaining parallel systems; and the technical feasibility of such an opt-out procedure). If these factors do not suggest the provision of an opt-out mechanism, the mode framework invites organisations to 'consider providing modes of recourse to the consumer, such as providing a channel for reviewing the decision' (ibid.).

Overall, the model framework is based on a number of reference ethical principles, whch the PDPC distilled from various sources, including the IEEE, the FATML, the OECD and the AI HLEG Ethics Guidelines for Trustworthy AI.

f. United Kingdom

Guide on using AI in the public sector

In collaboration with the Government Digital Service, in August 2019, the UK Office for Artificial Intelligence published a 'Guide on using AI in the public sector' (UK Government 2019). This is a collection of guidance documents, including how to assess whether the use of AI will help an administration to meet user needs, and how the public sector can best implement AI ethically, fairly and safely. The key characteristics and excerpts of this collection of guidance are provided below:

- *Definition of AI:* AI is defined as a research field spanning philosophy, logic, statistics, computer science, mathematics, neuroscience, linguistics, cognitive psychology and economics, and as 'the use of digital technology to create systems capable of performing tasks commonly thought to require intelligence'. The UK guidance primarily discusses machine learning but does not provide a technology-specific definition of AI. Machine learning is defined as a subset of AI and its most widely used form, and 'refers to the development of digital systems that improve their performance on a given task over time through experience' (ibid.).

could, to varying extents, be incorporated as part of the organisation's AI deployment process. Other measures, such as reproducibility, traceability and auditability, are more resource-intensive and may be relevant for specific features or in certain scenarios.' (ibid.)

- *Key factors to consider* in the development of AI include data quality (accuracy, completeness, uniqueness, timeliness, validity, sufficiency, relevancy, representativeness, consistency), fairness, accountability, privacy, explainability and transparency, costs. Institutions are invited to consider how much it will cost to build, run and maintain an AI infrastructure, train and educate staff and if the work to install AI may outweigh any potential savings (ibid.).

Table 7 - UK guidance on choosing the most appropriate machine learning technique

Machine learning technique	Description	Examples of machine learning technique
Classification	Learns the characteristics of a given category, allowing the model to classify unknown data points into existing categories	• Deciding if a consignment of goods undergoes border inspection • Deciding if an email is spam or not
Regression	Predicts a value for an unknown data point	• Predicting the market value of a house from information such as its size, location, or age • Forecasting the concentrations of air pollutants in cities
Clustering	Identifies groups of similar data points in a dataset	• Grouping retail customers to find subgroups with specific spending habits • Clustering smart-meter data to identify groups of electrical appliances and generate itemised electricity bills
Dimensionality reduction or manifold learning	Narrows the data to the most relevant variables to make models more accurate, or to make it possible to visualise the data	• Used by data scientists when evaluating and developing other types of machine learning algorithms
Ranking	Trains a model to rank new data based on previously seen lists	• Returning pages by order of relevance when a user searches a website

Source: UK Office for Artificial Intelligence (2019)

- Information Commissioner's Office Guidance on AI Auditing Framework

Following public consultation in February 2020, the Information Commissioner's Office (ICO) published its Guidance on AI Auditing Framework (ICO, 2020) covering best practice in the development and deployment of AI systems for ensuring compliance with data protection laws. The Guidance offers organisations a self-regulatory framework for assessing data protection

risks associated with the use of AI systems and makes recommendations on the best technical and organisational measures for mitigating those risks.

- *Voluntary framework:* While the Guidance is a voluntary framework, the ICO will use it as a methodological toolkit for its enforcement activities. As most organisations using AI technologies need to carry out mandatory Data Protection Impact Assessments (DPIA), the ICO recommends that companies align their DPIA processes with the requirements set out in the Guidance.
- *Addressees:* The Guidance is addressed to developers, designers, and deployers of AI systems. The ICO stresses that compliance specialists and technology experts should be actively involved throughout the lifecycle of AI systems in order to *meaningfully* address the data protection risks of AI technologies.

The four key themes of the Guidance are:

- *Accountability and governance*. The ICO highlights the importance of implementing holistic AI governance and risk management mechanisms in organisations. This means that even senior managers should be able to demonstrate risk mitigation strategies and justify the choices made in the organisation. Organisations should not simply delegate data protection to Data Protection Officers and diffuse responsibility.
    - DPIA: In addition to the standard elements of a DPIA (as required by the GDPR), the ICO recommends AI-specific components that organisations should include, such as:
        - A description of the degree of human involvement in the AI system's decision-making process;
        - Appropriate methods to describe data processing and the statistical accuracy of AI systems;
        - An evaluation of the proportionality and reasonableness of replacing human decision-making with AI by describing and documenting the trade-offs that are made (e.g. between statistical accuracy and data minimisation).
- *Lawfulness, fairness, and transparency*. The Guidance emphasises the need to find the appropriate legal basis for data processing at the different stages of AI systems. The Guidance notes that monitoring the processing of personal data at all stages is important because in some cases AI models learn to process special category data even when that was not the purpose of the original model. To maintain the principle of fairness and avoid bias, training data should be representative, balanced and ensure the highest possible level of statistical accuracy. Initially, organisations could use AI and human decision-making systems simultaneously to identify and flag limits of accuracy and bias. The ICO refers to its guidance report on transparency – explain (ICO, n.d.).
- *Security and data minimisation*. The Guidance lists the types of attacks and security breaches to which AI systems are vulnerable in the different phases of development and deployment, and recommends practical security measures to mitigate the risks of attacks (e.g. assessing security via penetration testing and applying external security certifications). The ICO recommends that security measures are applied in proportion to the likelihood and severity of the potential risk to individuals. While data minimisation is challenging in the context of AI systems, the Guidance stresses that organisations should only store and process data that is necessary and relevant. Organisations are

invited to remove irrelevant features from datasets and to minimise the data risk by applying formats that are less readable for humans.

- *Individual rights in AI systems*. The ICO provides suggestions for ensuring that individuals can effectively exercise their rights relating to their data. Some of the suggestions include:
    - Depending on the pre-processing methods and based on what the personal data are used for (e.g. training data), it may be difficult to respond to requests for access, rectification, and erasure of data. The ICO warns that stripping data of personal identifiers does not remove it from data protection obligations. In some cases, whole data models need to be erased.
    - Data portability might not apply to AI outputs, as those commonly constitute inferred data.
    - While individuals have the right to be informed where their data is used as training data for AI systems, it might be excessively difficult to inform persons directly where the dataset has been stripped of personal identifier features. The ICO therefore recommends that organisations provide public information on the source of the used data.

g. United States

US draft Guidance for Regulation of Artificial Intelligence Applications

In January 2020, the US Office of Management and Budget (OMB) published a request for comments on a 'Draft Memorandum to the Heads of Executive Departments and Agencies, "Guidance for Regulation of Artificial Intelligence Applications"' (Draft Memo, Vought 2020). The Draft Memo reflects the requirements of Executive Order 13859, 'Maintaining American Leadership in Artificial Intelligence' (the Executive Order), which called on the OMB to issue a memorandum that would '(i) inform the development of regulatory and non-regulatory approaches by such agencies regarding technologies and industrial sectors that are either empowered or enabled by AI, and that advance American innovation while upholding civil liberties, privacy and American values; and (ii) consider ways to reduce barriers to the use of AI technologies in order to promote their innovative application, while protecting civil liberties, privacy, American values, and US economic and national security.' (Exec. Order No. 13,859, 84 Fed. Reg. 3967, § 6(a); Van Demark 2020)

The Draft Memo built on the Executive Order and advocates a risk-based approach, stating that 'the magnitude and nature of the consequences should an AI tool fail, or for that matter succeed, can help [to] inform the level and type of regulatory effort that is appropriate to identify and mitigate risks' (Vought 2020). It adds that agencies should consider the degree and nature of the risks posed by various activities within their jurisdiction, where possible avoiding 'hazard based and unnecessarily precautionary approaches to regulation that could unjustifiably inhibit innovation' (LAIP, 2020).

Several aspects of the draft guidance are relevant here. The following text provides a summary of the most important excerpts from the draft memo (Vought 2020):

- *Definition of AI:* The definition adopted follows the one was given in Section 238(g) of the *John S. McCain National Defence Authorisation Act* for Fiscal Year 2019, Pub. L. No. 115- 232, 132 Stat. 1636, 1695 (13 August 2018) (codified at 10 U.S.C. § 2358, note), which defined AI to include the following (see Vought 2020):

- Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to datasets.
- An artificial system developed in computer software, physical hardware, or another context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.
- An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
- A set of techniques, including machine learning, that is designed to approximate a cognitive task.
- An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision-making, and acting.

The guidance focuses on 'narrow' (weak) AI that 'goes beyond advanced conventional computing to learn and perform domain-specific or specialised tasks by extracting information from datasets, or other structured or unstructured sources of information' (ibid.).

Risk assessment and management: The document acknowledges that 'When humans delegate decision-making and other functions to AI applications, there is a risk that AI's pursuit of its defined goals may diverge from the underlying or original human intent and cause unintended consequences - including those that negatively impact privacy, civil rights, civil liberties, confidentiality, security, and safety' (ibid.).

- Agencies should consider the risks of inadequate protection to algorithms and data throughout the design, development, deployment, and operation of an AI system, given the level of sensitivity of the algorithms and data. This includes the assessment of 'possible anti-competitive effects that favor incumbents at the expense of new market entrants, competitors, or up-stream or down-stream business partners'. The management of risks created by AI applications should be appropriate to and commensurate with the degree of risk that an agency determines in its assessment. Agencies are invited to adopt a 'tiered approach', in which the 'degree of risk and consequences of both success and failure of the technology determines the regulatory approach, including the option of not regulating' (ibid.).
  - For AI applications that pose *lower risks*, 'agencies can rely on less stringent and burdensome regulatory approaches - or non-regulatory approaches - such as requiring information disclosures or consumer education'.
  - For *higher risk* AI applications, 'agencies should consider the impact to the individual, the environments in which they will be deployed, the necessity or availability of redundant or back-up systems, the system architecture or capability control methods available when an AI application makes an error or fails, and how those errors and failures can be detected and remediated'.
- *Avoiding prescriptive regulation:* The draft guidance observes that 'Rigid, design-based regulations that attempt to prescribe the technical specifications of AI applications will in most cases be impractical and ineffective, given the anticipated pace with which AI will evolve and the resulting need for agencies to react to new information and evidence', and that 'Targeted agency conformity assessment schemes, to protect health and safety, privacy, and other values, will be essential to a successful, and flexible, performance-based approach'. Among the non-regulatory approaches considered, the document describes sector-specific policy guidance or

frameworks, pilot programmes and experiments, and voluntary consensus standards (ibid.).

- *Conformity assessment standards:* 'Federal agencies must use voluntary consensus standards in place of government-unique standards in their procurement and regulatory activities, except where inconsistent with law or otherwise impractical'. They are also invited to rely on the guidance in NIST publications to understand conformity assessment concepts and to use conformity assessment effectively and efficiently that meets agency requirements (ibid.).

New Jersey's Algorithmic Accountability Act

New Jersey's Algorithmic Accountability Act (NJAAA) was introduced on 20 May 2019, shortly after the federal Algorithmic Accountability Act was introduced in the US Congress. The proposal closely resembles the federal proposal and requires certain businesses and other 'covered entities' to carry out impact assessments on high-risk automated decision systems (ADS). Impact assessments need to be submitted to the Director of the Division of Consumer Affairs in the Department of Law and Public Safety (the Director). If adopted, the NJAAA will be a state-level binding policy framework.

- *Definition of high-risk ADS:* An ADS is considered high-risk if it (i) poses a 'significant risk' to the privacy or security of personal data; (ii) risks producing 'inaccurate, unfair, biased, or discriminatory decisions impacting consumers'; (iii) makes decisions based on the extensive analysis of sensitive aspects of consumers' lives (e.g. health and economic situation) that impacts them legally or otherwise; (iv) involves personally identifiable information of a significant number of consumers or it systematically monitors public spaces (New Jersey Algorithmic Accountability Act, A.B. 5430, 218th Leg., 2019 Reg. Sess., N.J. 2019).
- *Covered entities:* The addressees of the NJAAA are corporations, associations, organisations, and other legal entities that either (i) generate at least USD 50,000,000 per year; (ii) possess or control the personal data of one million New Jersey consumers or telecommunication devices; or (iii) are data brokers. The NJAAA does not apply to state or federal agencies.
- *Automated Decision System Impact Assessment (ADSIA):* ADSIA constitutes the evaluation of the ADS throughout its lifecycle and the assessment of its impact on 'accuracy, fairness, bias, discrimination, privacy, and security' (ibid.). Elements of the ADSIA include:
    - Cost-benefit analysis of the ADS in light of its purpose and other factors, such as:
        - data minimisation practices;
        - duration for which consumers' personal data are stored;
        - information available to consumers on the ADS;
        - consumers' access to the results of the ADM process and possibility to object to them;
        - security threats to consumers' personal data in the information system;
        - assessment of the risks posed by ADS that may result in inaccurate, unfair, biased, or discriminatory decisions impacting consumers.

- A detailed description of best practices (including technological and physical) used to minimise risks identified during the analysis;
- Cooperation with independent third parties (e.g. auditors and technology experts) for conducting the impact assessment;
- Obligation to record any bias and security threats to consumers' personal data.

- *Civil law consequences of the NJAAA:* An agreement between a covered entity and a consumer which does not comply with the NJAAA is void and unenforceable. If the Director concludes, based on the ADSIA, that the ADS poses a threat to consumers or negatively impacts consumers, the Attorney General of the State can initiate civil action on behalf of affected consumers to obtain compensation. Unlawful behaviour contrary to the NJAAA constitutes a violation of the Consumer Fraud Act, which may imply a penalty of USD 20,000.

2. Other proposed policy initiatives on AI

In recent months, governments, international and civil society organisations have formulated proposals to establish regulatory or self-regulatory frameworks to ensure responsible development of AI. The rising importance of AI in the global order is inevitably reflected in lively discussions among international organisations. Early initiatives were adopted at regional level (e.g. the Declaration on AI in the **Nordic-Baltic Region**, issued in May 2018 by Denmark, Estonia, Finland, the Faroe Islands, Iceland, Latvia, Lithuania, Norway, Sweden and the Åland Islands) and within **UNESCO** (Report of COMEST on Robotics Ethics, World Commission on the Ethics of Scientific Knowledge and Technology, September, 2017).

Based on the work of the European Commission and AI HLEG, the **OECD** adopted principles on AI, stressing the need for innovative and trustworthy AI and the need to respect human rights and democratic values. The OECD principles were presented in the form of an OECD Council Recommendation on Artificial Intelligence[60]. They involve all OECD members and Argentina, Brazil, Colombia, Costa Rica, Peru and Romania. The OECD AI Principles complement other OECD standards in areas like digital security, privacy, risks and responsible commercial activities. Moreover, in June 2019, the **G20** adopted AI principles based on the OECD AI principles[61]. These developments have been accompanied by an equally important debate in the UN, in particular within the **International Telecommunications Union**, which has promoted the 'AI for good' platforms, linking AI to the SDGs. This process is flanked by the 'AI for Humanity' idea that emerged from the elaboration of Mission Villani for the French government, and led to the creation of a movement that culminated in a **Global Forum on AI for Humanity (GFAIH)** in October 2019 in Paris under the auspices of the French government. The GFAIH was intended as the formal launch of the so-called Global Partnership on AI (GPAI) and will inform the future agenda of GPAI Working Groups.

The **GPAI** is worth highlighting, as it emerges from the agenda of the **G7**. Since 2018, Canada and France (and to some extent, Japan) have proposed the creation of an International Partnership on AI (IPAI). In December 2018, at the G7 Multistakeholder Conference on Artificial Intelligence, and following the Canada-France Statement on Artificial Intelligence (Government of Canada, 2018), Prime Minister Justin Trudeau, Minister Bains and Cédric O, France's Secretary of State for the Digital Sector, announced a mandate for the IPAI[62]. In May 2019, the Declaration and organisational structure of the IPAI were made public at the

[60] https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[61] See https://www.oecd.org/going-digital/ai/principles/ and https://www.mofa.go.jp/files/000486596.pdf

[62] https://pm.gc.ca/en/news/backgrounders/2018/12/06/mandate-international-panel-artificial-intelligence

end of the informal meeting of G7 Digital Ministers[63]. However, the IPAI faced swift opposition from the US and China (and also experienced funding issues). After a failed attempt to create IPAI in the G7 meeting in Biarritz, it was transformed into GPAI but has yet to be fully endorsed by the US and seems to suffer from rivalry between the US and China (Simonite, 2020). Important projects and international cooperation are happening also in other forums, such as the Organisation for Security and Cooperation in Europe (**OSCE)** (especially on the side of freedom of expression), the **Council of Europe**, and many more.

Table 8 - Ethical principles identified in existing AI guidelines

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

Source: Jobin et al. (2019)

Global governance is increasingly characterised by the growing role of non-state actors, and the AI landscape is no exception. While a full description of the numerous initiatives that have emerged in the AI domain would go beyond the scope of this study, it is important to mention the role of the private sector, in particular multi-stakeholder initiatives, in shaping the global landscape of human-centric AI (AlgorithmWatch n.d.). Notable examples include the US

[63] https://www.canada.ca/en/innovation-science-economic-development/news/2019/05/declaration-of-the-international-panel-on-artificial-intelligence.html

Association for Computing Machinery (USACM) 'Statement on Algorithmic Transparency and Accountability (2017), China's AI Industry Alliance's Joint Pledge on Artificial Intelligence Industry Self-Discipline (2019), DeepMind's Ethics & Society Principles (n.d.), FATML's Principles for Accountable Algorithms and a Social Impact Statement for Algorithms (2016), the Asilomar AI Principles (2017), Information Technology Industry Council (ITI) AI Policy Principles (2017), the Japanese Society for Artificial Intelligence Ethical Guidelines (2017), the Tenets of the Partnership on AI to Benefit People and Society (n.d.), the 12 Universal Guidelines for the Development of AI issued by the Public Voice Coalition (2018), a group of NGOs and representatives assembled by EPIC, and the UNI Global Union Top 10 principles for ethical artificial intelligence (2017).

Jobin et al (2019) surveyed existing ethical guidelines in the field of AI and found as many as 84 documents containing ethical principles or guidelines, most of which were released after 2016 and produced by private companies and government agencies. Table 9 shows the relative diffusion of individual ethical principles in those documents, with transparency and justice and fairness featuring most strongly.

Two international standards bodies are currently developing AI standards (see Cihon 2019):

- A common effort between **ISO and IEC** to coordinate the creation of digital technology standards. ISO and IEC founded a committee (JTC 1) in 1987, which has published some 3,000 standards, which were adopted by leading multinational corporations.
- The **IEEE** Standard Association, an engineers' professional organisation, creates process standards areas such as software engineering and management and autonomous systems design. Its AI standardisation processes are part of a IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. The **IEEE** hosted the development of ethically aligned design standards and 'applicable laws and regulations' more broadly, promoting a 'Vision for prioritising human well-being with autonomous and intelligent systems'[64]. In 2019, the IEEE released its 'Ethical aspects of autonomous and intelligent systems' (IEEE, 2019).

Table 10 summarises the most important standardisation processes in the AI domain.

[64] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf

Table 9 - International landscape of AI standards

Network - Product	Network - Process
• Foundational Standards: Concepts and terminology (SC 42 WD 22989), Framework for Artificial Intelligence Systems Using Machine Learning (SC 42 WD 23053) • Transparency of Autonomous Systems (defining levels of transparency for measurement) (IEEE P7001) • Personalized AI agent specification (IEEE P7006) • Ontologies at different levels of abstraction for ethical design (IEEE P7007) • Wellbeing metrics for ethical AI (IEEE P7010) • Machine Readable Personal Privacy Terms (IEEE P7012) • Benchmarking Accuracy of Facial Recognition systems (IEEE P7013)	• Model Process for Addressing Ethical Concerns During System Design (IEEE P7000) • Data Privacy Process (IEEE P7002) • Methodologies to address algorithmic bias in the development of AI systems (IEEE P7003). • Process of Identifying and Rating the Trustworthiness of News Sources (IEEE P7011)
Enforced - Product	**Enforced - Process**
• Certification for products and services in transparency, accountability, and algorithmic bias in systems (IEEE ECPAIS) • Fail-safe design for AI systems (IEEE P7009)	• Certification framework for child/student data governance (IEEE P7004) • Certification framework for employer data governance procedures based on GDPR (IEEE P7005) • Ethically Driven AI Nudging methodologies (IEEE P7008)

Source: Cihon (2019)

In 2017, JTC 1 founded Sub-Committee (SC) 42 to focus on standards for AI systems. The Secretariat is situated in the US. The main objectives of the Committee are to serve as the focus and proponent for JTC 1's standardisation programme on AI and to provide guidance to JTC 1, IEC, and ISO committees on the development of AI applications. In January 2020, SC 42 had 29 participating members and 13 observing members. SC 42 has published three standards, including two technical reports, with many more in development up until today (see Standards Australia 2020 and Table 11).

Table 10 - Standards under development in ISO/IEC JTC 1/SC 42

Project	Focus area
ISO/IEC TR 20547-2:2018	Information technology — Big data reference architecture — Part 2: Use cases and derived requirements
ISO/IEC TR 20547-5:2018	Information technology — Big data reference architecture — Part 5: Standards roadmap
ISO/IEC AWI 38507	Information technology — Governance of IT — Governance implications of the use of AI by organizations
ISO/IEC CD 22989	Artificial intelligence — Concepts and terminology
ISO/IEC CD 23053	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
ISO/IEC CD TR 20547-1	Information technology — Big data reference architecture — Part 1: Framework and application process
ISO/IEC AWI 24668	Information technology — Artificial intelligence —Process management framework for Big data analytics
ISO/IEC FDIS 20547-3	Information technology — Big data reference architecture — Part 3: Reference architecture
ISO/IEC 20546:2019	Information technology — Big data — Overview and vocabulary
ISO/IEC NP 24029-2	Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Formal methods methodology
ISO/IEC AWI TR 24368	Information technology — Artificial intelligence — Overview of ethical and societal concerns
ISO/IEC CD TR 24029-1	Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
ISO/IEC PDTR 24028	Information technology — Artificial Intelligence (AI — Overview of trustworthiness in Artificial Intelligence
ISO/IEC NP TR 24027	Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making
ISO/IEC AWI 23894	Information Technology — Artificial Intelligence — Risk Management
ISO/IEC CD TR 24030	Information technology — Artificial Intelligence (AI) — Use cases
ISO/IEC NP TS 4213	Information technology — Artificial Intelligence — Assessment of classification performance for machine learning models
ISO/IEC AWI TR 24372	Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems

Source: Standards Australia (2020)

Finally, several research centres and think tanks around the world are proposing policy frameworks. The **AI Now Institute** presented a practical framework for evaluating ADS deployed by public agencies (Reisman et al., 2018). Their report addresses both public agencies and affected communities. Firstly, it invites agencies to carry out AIA in the procurement process of ADS. Secondly, the framework allows experts and the public to gain insight and judge whether a given AI technology meets accepted standards of fairness, transparency, and security.

AI Now aims to 'end the use of unaudited black box systems' and facilitate informed policy debate and public engagement (Campolo et al., 2017). The public agency AIA model is not intended to be a comprehensive regulatory framework but, rather, serves several policy goals, including creating an accountability framework for public agencies, increasing public agencies' expertise, and building capacity for the internal evaluation of ADS.

The steps of the AIA process are:

- *Pre-acquisition review:* The procuring agency allows the public to flag any concerns or comment on any ADS before the contract is concluded.
- *Initial agency disclosure requirements:* The agency will (i) publish its internal domain-specific definition of ADS; (ii) disclose to the public extensive information relating to the purpose, uses and implementation of each ADS; (iii) carry out a self-assessment to identify risks related to fairness, justice, bias and inaccuracy, and describe how it will address these problems; (iv) provide a plan for how external researchers will be able to review the system following deployment.
- *Comment period:* At this stage, the agency invites the public to comment on the initial agency disclosure, taking into account the evidence presented by researchers.
- *Due process challenge period:* In order to make sure that the concerns raised by the public are addressed, the AIA process gives the public the opportunity to challenge the agency's decision to deploy an ADS before an agency oversight body or court.
- *Renewing AIAs:* Agencies are required to renew AIAs regularly. This includes incorporating new research findings and having periodic comment and due process periods.

ANALYSIS OF THE RESULTS OF THE PUBLIC CONSULTATION ON THE EUROPEAN COMMISSION WHITE PAPER ON AI

The analysis of the open public consultation consists of two separate analyses: (1) The analysis of 18 free text questions from the questionnaire of the consultation on the White Paper on AI, with a total of 6,667 free text responses; and (2) the analysis of 408 position papers submitted to the public consultation.

(1) The analysis of the free text responses is available in an Excel file attached to this report. For each of the 18 questions, it contains an overview table ranking the most prominent arguments put forth by stakeholders, as well as a breakdown by stakeholder type. In addition, each aggregate analysis is accompanied by an explorable datasheet, which contains the raw data and the full response texts. The sheets also contain an overview of the methodology and a brief explanation of using the Excel report to explore stakeholder feedback. The Excel report represents task 4a in the Terms of Reference (ToR).

(2) The analysis of the 408 position papers is summarised in this report, with a detailed analysis provided in a second Excel file. That Excel report contains additional aggregate data, as well as a user guide to using the underlying raw data to dive deeper into stakeholder feedback. That Excel report also contains a brief summary of each position paper. Together, these two reports represent task 4b-d ToR.[65]

A significant part of the analysis for this deliverable is thus available in the two Excel reports. The following sections here provide a summary of the analysis of the position papers only, specifically the findings on: main arguments; the definition of AI; the costs of AI regulation; the institutional governance of AI; and the regulatory requirements.

1. Main arguments in position papers

This section presents the main arguments put forward by stakeholders. The classification of main arguments was created by extracting up to three central points from each position paper, without predefined topics. As the four subsequent sections will look closely at four predefined topics (the definition of AI; costs; institutional governance; and regulatory requirements), they are excluded from this first overview section.

All numbers should be read with an ‘at least’ qualifier (‘at least 74 stakeholders believe that ...’) because only up to three main arguments were extracted for each position paper. Stakeholders may share other positions, but only their three central points were considered for this analysis. Table 12 provides an overview of the different types of stakeholders who submitted a position paper to the consultation.

[65] See Annex 2 for methodological details.

Table 11 - Number of position papers by stakeholder type

Stakeholder type	Count of position papers	% of position papers
Unspecified	94	23.0%
NGO	72	17.6%
Business Association	60	14.7%
Company (Large)	53	13.0%
Academic/Research Institution	49	12.0%
EU Citizen	24	5.9%
Company (SME)	21	5.1%
Public authority	19	4.7%
Trade Union	8	2.0%
Non-EU Citizen	6	1.5%
Consumer Organisation	2	0.5%
Grand Total	408	100%

Source: Public Consultation

a. Key findings

- The most important point for many respondents is the definition of 'high-risk'. A large group of stakeholders believe that the definition of high-risk is unclear or needs improvement (at least 18% of all stakeholders, 74 out of 408). Many believe that the binary classification in high vs. low is too simplified and some propose introducing more levels of risk. Some believe that the definition is too broad, while others believe that it is too narrow.
- Some stakeholders propose alternative approaches to defining 'high-risk' with more risk levels: at least six position papers suggest following the GDEC's gradual approach, with five risk levels to create a differentiated scheme of risks. Others suggest the adoption of risk matrices, which combine the intensity of potential harm with the level of human implication/control in the AI decision. The probability of harm is another criterion repeatedly mentioned by stakeholders.
- Many position papers address the proposed two-step approach to determining 'high-risk' AI. At least 19 believe that the approach is inadequate, at least five argue against a sectoral approach, and many others put forward a diverse set of suggestions and criticisms.
- One notable suggestion for the risk assessment approach is to take into account all subjects affected by the AI application: multiple stakeholders argue that collective as well as individual risks should be considered, as there are risks to society as a whole (e.g. democracy, environment, human rights).
- At least 52 stakeholders addressed the proposed voluntary labelling scheme. At least 21 position papers are sceptical of labelling, either because they believe that it will impose regulatory burdens (especially for SMEs) or because they are sceptical about its effectiveness. Some stakeholders argue that such a scheme is likely to confuse

consumers instead of building trust. By contrast, at least eight position papers are explicitly in favour, with many others providing a diverse set of comments.

- 52 address issues of liability, with most providing a diverse set of comments. Eight believe that existing rules are likely sufficient. At least six are sceptical about a strict liability scheme, noting that such a scheme is likely to stifle investment and innovation, and that soft measures like codes of conduct or guidance documents are better suited. At the same time, many more contributions to the public consultation from the entire range of stakeholder types express support for a risk-based approach with respect to liability for AI, suggesting that not only the producer but also other parties should be liable. Representatives of consumer interests stress the need for a reversal of the burden of proof.
- Many position papers underline the importance of fundamental rights and other ethical issues. The importance of fundamental rights in AI regulation is underlined by at least 42 position papers, six of which argue in favour of human rights impact assessments.
- Many respondents highlight ethical issues such as discrimination and bias (21), the importance of societal impacts (18), data protection (15), civil society involvement (nine) or human oversight (seven). The Excel file accompanying this report gives a broader view of the remaining principal arguments that were shared by stakeholders, albeit in smaller numbers.

Table 12 - Detailed arguments linked to the top five topics in the position papers

Arguments linked to top 5 topics	Count of Arguments
Risk definition	**85**
Definition of 'high-risk' is unclear/needs improvement	74
Risk definition (other comments)	11
Two step risk assessment approach	**73**
Risk assessment approach (other comments)	42
Risk assessment method is inappropriate	19
Against sectoral approach to "High-risk"	5
Regulation should apply to more than high risk	5
Consider that AI can change	2
Voluntary labelling	**52**
Sceptical of labelling (e.g. due to costs or lack of effectiveness)	21
Other comments	15
In favour of voluntary labelling	8
Sceptical of labelling being voluntary / self-labelling	4
More details needed	4
Liability	**52**
Other comments	35
Existing rules probably sufficient	8
Sceptical of strict liability / new regulatory burdens	6
New liability rules needed	3
Fundamental rights	**42**
Other comments on importance of FR in AI regulation	36
In favour of human rights impact assessment	6
Grand Total	**304**

Source: Public Consultation

b. Breakdown by stakeholder

- The impression that the definition of 'high-risk' needs to be clarified is shared by all stakeholder types.

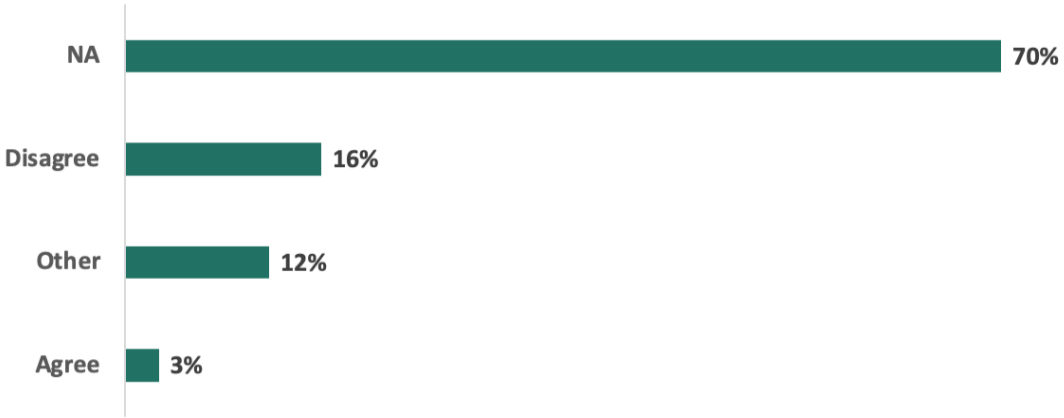
- The two-step risk assessment approach received most comments from business stakeholders. At least five business associations and large companies argue against the sectoral approach to determining high-risk and are more supportive of a contextual assessment. By contrast, two of the three SMEs that mentioned risk assessment expressly supported a sectoral approach.
- The voluntary labelling scheme also received most comments from business stakeholders: most of the business associations (at least 11) and SMEs (at least three) are sceptical about the idea, due to the likely costs incurred or its suspected lack of effectiveness. The position of the large companies that mentioned voluntary labelling is quite the opposite, with most (at least four) tending to be in favour.
- Many business associations and large companies think that existing rules on liability are already sufficient (at least seven) or they are sceptical about strict liability rules and possible regulatory burdens (at least five). This position is shared by almost none of the other stakeholder types.
- Fundamental rights issues are mostly emphasised by NGOs (at least 16), at least five of which are in favour of introducing a human rights/fundamental rights impact assessment for AI.

2. Definition - how to define AI?

This section contains the results of the analysis of the stakeholders' positions on the definition of AI in the White Paper. As the White Paper does not contain its own explicit definition, this analysis took the definition of the AI HLEG as a reference point, which includes systems that use symbolic rules or machine learning, but does not explicitly include simpler ADM systems.

Every position paper was analysed to determine whether and why stakeholders agree or disagree with this definition, or have other useful comments on the definition of AI.

Figure 14 - Stakeholders' positions on the definition of AI



a. Key findings

- The majority of position papers make no mention of the definition of AI (up to 70%, or 286 out of 408).
- A majority (15.7%) disagree with the position of the Commission (at least 64). At least 9.3% state that the definition is too broad (37), of which 2.7% say that AI should only include machine learning (11). Stakeholders highlight that an overly broad definition risks over-regulation and legal uncertainty, and is not specific enough to AI. At least

6.6% believe that the definition is too narrow (27), with 3.7% saying that it should also include ADM systems (15). Stakeholders here highlight that an overly narrow definition misses many dimensions that will build the future generation of AI.

- At least 2.7% of stakeholders agree with the Commission/AI HLEG definition of AI (11).
- At least 5.4% of position papers state that the Commission's definition is unclear and needs to be refined (22). To improve the definition, stakeholders propose: clarifying the extent to which the definition covers traditional software; distinguishing between different types of AI; looking at existing AI definitions from public and private organisations. Finally, at least 2.2% of stakeholders provide their own definition of AI (9).

Table 13 - Overview of stakeholder's positions on the definition of AI

Position on definition	Nb of position papers	% of position papers
⊞ NA	286	70.1%
⊟ Disagree	64	15.7%
Disagree - Too broad (other)	26	6.4%
Disagree - Too narrow (should also include ADM)	15	3.7%
Disagree - Too narrow (other)	12	2.9%
Disagree - Too broad (should be only ML)	11	2.7%
⊟ Other	47	11.5%
Definition is unclear, needs to be refined	22	5.4%
Provides own definition of AI	9	2.2%
Other comments on the definition of AI	8	2%
Align definition with OECD	4	1%
Suggesting modifications to the HLEG-AI definition	2	0.5%
Differentiate between different types of AI	2	0.5%
⊞ Agree	11	2.7%
Grand Total	408	100%

Source: Public Consultation

b. Breakdown by stakeholder type

- When the definition of AI was mentioned, the key difference between stakeholders concerned the scope of the definition. The majority of business stakeholders believe that the Commission's definition is too broad, with the trend strongest among business associations. By contrast, the majority of academic and NGO stakeholders believe that the Commission's definition is too narrow.
- At least 24 business stakeholders believe that the definition is too broad, while only five believe that it is too narrow and only four agree with it. Business stakeholders are also relatively numerous in saying that the definition is unclear/needs to be refined (at least 11).
- The majority of academic and NGO stakeholders believe that the Commission's definition is too narrow (at least six and eight, respectively), while only one academic and four NGO stakeholders believe that the definition is too broad.

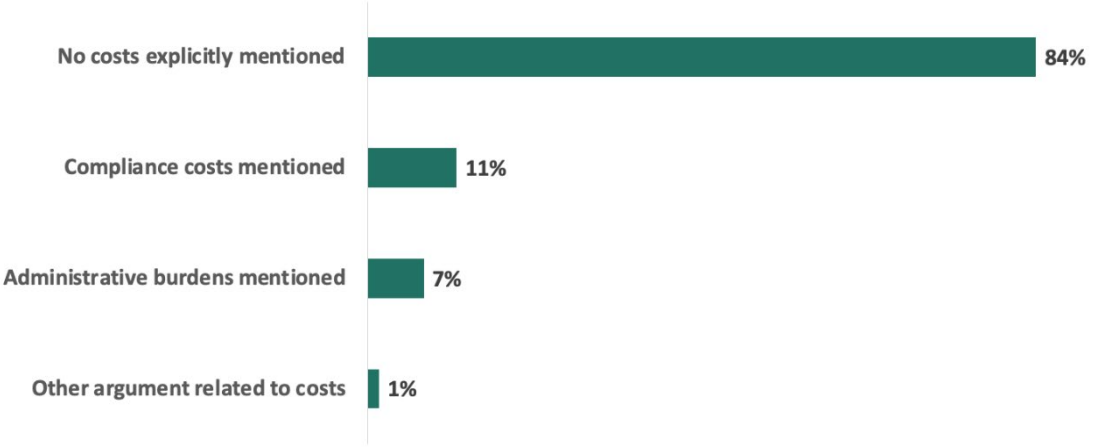
3. Costs - what costs could AI regulation create?

Costs imposed by new regulations are always a contentious topic. Some see costs imposed by regulation as an unnecessary burden to competitiveness and innovation, while others see them as a necessary by-product of organisations' compliance with political, economic or ethical objectives.

In order to better understand stakeholders' perspectives on the costs of AI regulation, every position paper was analysed for mentions of two main types of costs: (1) compliance costs, generally defined as any operational or capital expense faced by a company to comply with a regulatory requirement; and (2) administrative burdens, a subset of compliance costs, covering 'red tape' such as obligations to provide or store information.

a. Key findings

Figure 15 - Mention of costs in the submissions



- Up to 84% of stakeholders do not explicitly mention costs that could be imposed by a regulation on AI (344).
- At least 11% stakeholders (46) mention compliance costs in general, and at least 7% stakeholders (29) (also) mention administrative burdens in particular. Some stakeholders mention both types of costs.
- Some stakeholders warn against the costs incurred by a mandatory conformity assessment, especially for SMEs or companies operating in international markets. Some highlight that certain sectors are already subject to strict ex ante conformity controls (e.g. automotive sector) and warn against legislative duplication. Several stakeholders also see a strict liability regime as a potential regulatory burden, with some noting that a stricter regime can lead to higher insurance premiums.
- Some respondents put forward other arguments related to costs, such as the potential cost-saving effects of AI, the concept of 'regulatory sandboxes' as a means to reduce regulatory costs, or the environmental costs created by AI due to high energy consumption.

b. Breakdown by stakeholder type

- 17% of all types of business stakeholders mention compliance costs and 13% (also) mention administrative burdens, while up to 74% do not explicitly mention costs. Among business stakeholders, business associations most frequently mention costs. Of all mentions of costs by all stakeholders (75 in total), 56% come from business

stakeholders (42). One example of the business position on compliance costs is put forward by the US Chamber of Commerce:

'A new conformity assessment regime would likely serve as a significant bottleneck in the development and deployment of AI in the EU, as companies would need to win approval from regulators before deploying AI-enabled goods and services in the Single Market. Many innovative small and medium-sized enterprises that may have neither the time nor resources to undergo such a process will either avoid investing in perceived "high risk" areas or deploy their solutions abroad. The additional costs will reduce competition and choice in the Single Market for AI goods and services deemed as "high risk".'

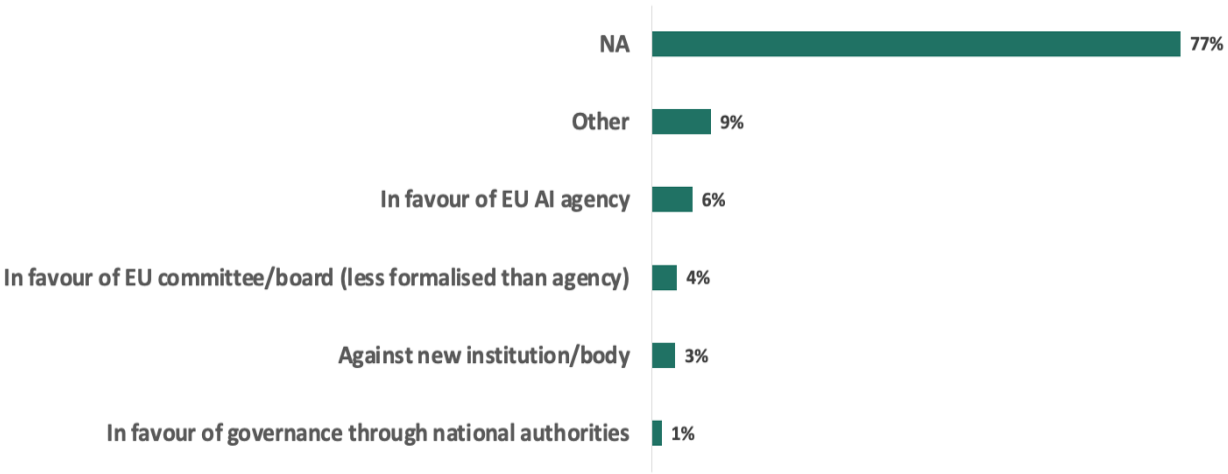
- Academic stakeholders mention costs more often than other types of stakeholders, although not very often overall. At least 13% of academic stakeholders mention compliance costs and 9% (also) mention administrative burdens, while 82% do not explicitly mention costs.
- Other stakeholders mention costs more rarely.

4. Governance - which institutions could oversee AI governance?

The institutional structure of AI governance is a key challenge for the European regulatory response to AI. Should AI governance be centralised in a new EU agency, for example, or decentralised in existing national authorities, or something in between? In order to better understand this issue, all position papers were analysed for their position on the European institutional governance of AI.

a. Key findings

Figure 16 - Stakeholders' positions on the governance of AI



- Most stakeholders (up to 77% or 314) did not address the institutional governance of AI.
- Of the 23% that addressed this issue, a majority of at least 10% of stakeholders are in favour a new EU-level institution, with at least 6% favouring some form of a new EU AI agency (24) and at least 4% a less formalised EU committee/board (15). At the same time, at least 3% stakeholders are against establishing a new institution (14), arguing that creating an additional layer of AI-specific regulators could be counterproductive and instead advocating for a thorough review of existing regulation

frameworks (e.g. lessons learned from data protection authorities dealing with GDPR) before creating a new AI-specific institution/body.

- At least 1% of stakeholders are in favour of governance through national institutions (six) and another 1% are in favour of governance through existing competent authorities (five) (without specifying whether these would be at EU or national level).
- Stakeholders also mention other ideas, such as the importance of cooperation between national and/or EU bodies (sevens); multi-stakeholder governance involving civil society and private actors (six); or sectoral governance (four).

Table 14 - Stakeholders' positions on the institutional governance of AI

Position on Institutional Governance of AI	Nb positions
NA	**314**
Other	**35**
Cooperation between the responsible national and/or EU authorities is important	7
In favour of multi-stakeholder governance	6
In favour of governance through existing competent authorities	5
In favour of sectoral governance	4
Clarity and effectiveness is important	4
In favour of democratic and human rights-centered governance	3
Other comments on European AI centres for R&D	3
In favour of global treaty on AI	1
In favour of Global AI Governance Agency	1
In favour of only minimal harmonisation between national and EU bodies	1
In favour of EU AI agency	**24**
In favour of EU committee/board (less formalised than agency)	**15**
Against new institution/body	**14**
In favour of governance through national authorities	**6**
Grand Total	**408**

b. Breakdown by stakeholder type

- While only 32% of academic stakeholders mention the issue, they tend to be in favour of an EU AI agency (at least 10%), but many provide a diverse set of other arguments.
- 24% of large companies and business associations have a position on the issue, while SMEs scarcely mention it. All business stakeholders tend to be more sceptical of formal institutionalisation: 8% of business associations and 4% of large companies are against a new institution, 5% of associations and 2% of large companies are in favour of a less formalised committee/board, and the others share other more specific positions.
- Most trade unions and EU/non-EU citizens do not have a position on the issue, while those that do are mostly in favour of an EU AI agency (25% of trade unions and 17% of EU/non-EU citizens). These percentages are unreliable, however, due to the low numbers of respondents with a position on the issue.
- More details on stakeholders' positions can be found in the Excel report. One example is the position from OpenAI:

'We believe that a cross-country governance structure can help the Commission [to] address the transnational nature of AI and its associated governance challenges. Such a structure would benefit from a permanent secretariat, along with an assembled committee of experts. By having a permanent secretariat, it would be possible to fund and conduct continuous measurement, assessment, and "spot check" activities, which would provide valuable information for EU citizens, elected officials, and the assembled committee of experts. Possible members of the secretariat community could include institutions like the OECD's AI Policy Observatory, with which the Commission is already collaborating via the Joint Research Centre. This governance structure could include permanent members from multiple European countries to reflect both regional and sub-national concerns.'

5. Regulatory requirements for 'high-risk' AI

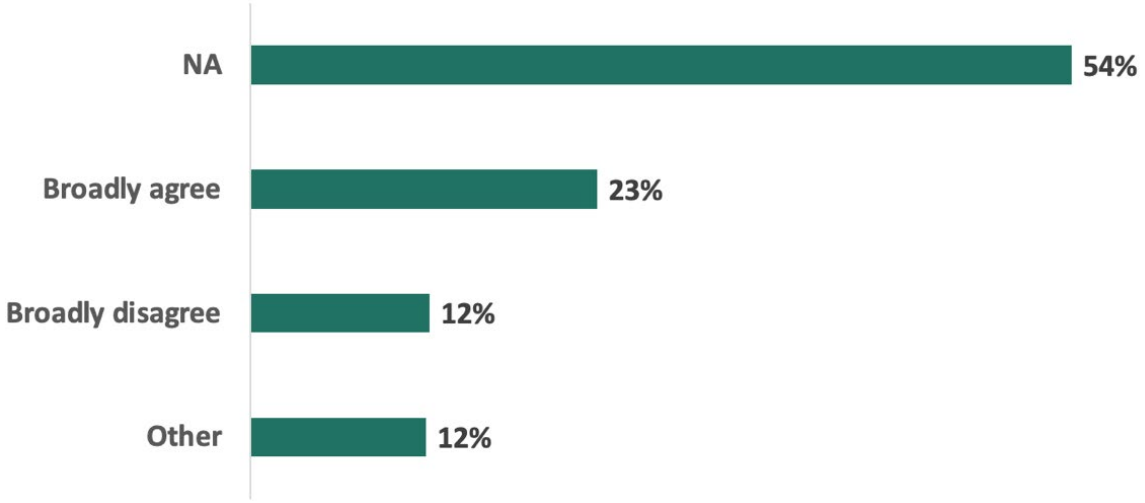
A cornerstone of any regulation is the mandatory requirements tailored for high-risk AI: human oversight, training data, data and record-keeping, information provision, robustness and accuracy, and specific additional requirements for certain AI applications, such as RBI.

In order to understand the position of the different stakeholders on the proposed regulatory requirements, the study team analysed the mentions of the requirements, as well as the general position on the requirements ('Broadly agree'; 'Broadly disagree'; 'NA'; 'Other').

Specific attention was also paid to RBI, with every position on this specific group of controversial technologies analysed.

a. Key findings

Figure 17 - Stakeholders' positions on regulatory requirements for high-risk AI



While more than half of the position papers do not mention regulatory requirements (54%), many (at least 23%) generally agree with the White Paper's approach to regulatory requirements for high-risk AI. At least 12% generally disagree, and some stakeholders express other opinions (12%).

Of the 12% of stakeholders who express another opinion (47), at least 1.7% argue that no new AI requirements are needed (seven), while another 1.7% ask for additional requirements (e.g. on intellectual property or AI design) to be considered (seven). Other comments highlight

that the requirements must not stifle innovation (six) or that they needed to be more clearly defined (three).

Table 15 - Stakeholders' positions on regulatory requirements for high-risk AI

Positions on Regulatory Requirements	Nb positions
⊞ **NA**	**220**
⊞ **Broadly agree**	**93**
⊞ **Broadly disagree**	**48**
⊟ **Other**	**47**
Requirements should be applied proportionately to risk and context	8
Additional requirements should be considered	7
New AI regulation/requirements are not needed	7
Definition of 'high-risk' is unclear/needs improvement	6
Requirements must not stifle innovation	6
Requirements should apply to all AI applications (not only 'high-risk')	5
Requirements should be reviewed as AI's risk evolves	3
Requirements must be more clearly defined	3
Other comments	2
Grand Total	**408**

'Human oversight' is the most frequently mentioned requirement (109 mentions), followed by 'training data' (97), 'data and record-keeping' (94), 'information provision' (78), and 'robustness and accuracy' (66).

At least 24% of the stakeholders specifically mention RBI (93). At least 4.7% argue for a ban on RBI in public spaces (19), and at least 1.7% for a moratorium (seven). At least 4.7% are in favour of conditioning its use through tight regulation and adequate safeguards (19).

Table 16 - Stakeholders' positions on RBI requirements

Positions on remote biometric identification	Nb positions
RBI should only be allowed with tight regulation and sufficient safeguards	19
In favour of a ban of RBI in public spaces	19
RBI should always be considered 'high-risk'	12
Definition/use/regulation of RBI should be clarified	11
GDPR can be used to regulate RBI	9
A broader debate on RBI is necessary	8
Other comments	8
In favour of a moratorium on RBI	7
Grand Total	**93**

b. Breakdown by stakeholder type

- Many business associations (73%) and large companies (59%) take a stance on regulatory requirements, while the other stakeholder types, including SMEs, do so less frequently. Business stakeholders tend to broadly agree with the Commission's approach (at least 31%). Those who express other opinions primarily highlight that new rules/requirements are not needed (3.7%), or that requirements should be proportionate (2.2%).

- Only 39% of academic stakeholders mention regulatory requirements (19). When they do, they tend to be in favour (22%) or to express other opinions (10%). The positioning of NGOs is similar: while only 38% mention regulatory requirements, those who do are generally in favour (21%).
- Almost half of the stakeholders who are in favour of a ban on RBI in public spaces are NGOs. This contrasts with the 34 business stakeholders who mentioned RBI, only one of which is in favour of a ban.
- A moratorium on RBI is more popular among academic stakeholders, with 33% of research institutions in favour of such a moratorium until clear and safe guidelines are issued by the EU (four).

ASSESSMENT OF THE COMPLIANCE COSTS GENERATED BY THE PROPOSED REGULATION ON AI

1. Methodology

The cost assessment here relies on the Standard Cost Model, a widely known methodology for the assessment of administrative burdens. Originally developed in the Netherlands, it was later adopted and expanded to the broader category of direct compliance costs by several countries around the world, including almost all EU Member States and the European Commission in its Better Regulation Toolbox (Renda et al., 2013; European Commission, 2015; Renda et al., 2019). A specific version of the Model is used here, as proposed by the German Federal Government, which has the additional advantage of featuring standardised tables with time estimates per administrative activity and level of complexity.

The cost estimation is built on time expenditure on activities induced by the new requirements under the proposed regulation. The assessment is based on cost estimates of an average AI *unit* of an average firm. This is then used to divide the total AI market in Europe into a number of AI units. The number of AI units is then multiplied by the cost per unit to reach an estimate of the compliance costs generated by the proposed regulation. The overall approach is simplified below:

$$Total\ compliance\ cost = \left(\frac{Value\ of\ the\ European\ AI\ market}{Value\ of\ an\ AI\ unit}\right) \times Cost\ per\ AI\ unit$$

Given the uncertainty of the evolution of the AI market, the final result is given within an upper and lower bound, based on two different estimations of the evolution of the AI market[66].

Two workshops were organised in order to close information gaps. Stakeholders from key businesses were invited to discuss the study team's estimates for compliance costs. Similarly, accreditation bodies and standardisation organisations were invited to another workshop to discuss the team's estimates on conformity costs.

According to different sources, a customized AI may cost USD 100,000 to USD 300,000[67]. This report takes USD 200,000 or EUR 170,000 as the reference value of an AI unit (i.e. a unit of value of EUR 170,000 of an AI system). An AI system thus consists of less than one unit or multiple units. The cost estimate per AI unit will be multiplied by the estimated number of AI units developed in a year in order to obtain an estimate of the costs of total compliance to the economy. The advantage of this approach is that the cost will be linear to the total AI investment, while cost estimates of each requirement do not require further analysis by level of sophistication. The use of AI unit as the unit of analysis will help to capture the fact that a more complex AI will incur higher costs of compliance and conformity assessment.

[66] Lower and upper bounds are included for the projection to the population, but not for all cost estimates. As a series of assumptions had to be made, including bounds for each would push the uncertainty to a very high level.

[67] For AI costs, see https://www.webfx.com/internet-marketing/ai-pricing.html, https://azati.ai/how-much-does-it-cost-to-utilize-machine-learning-artificial-intelligence/ and https://www.quytech.com/blog/ai-app-development-cost/

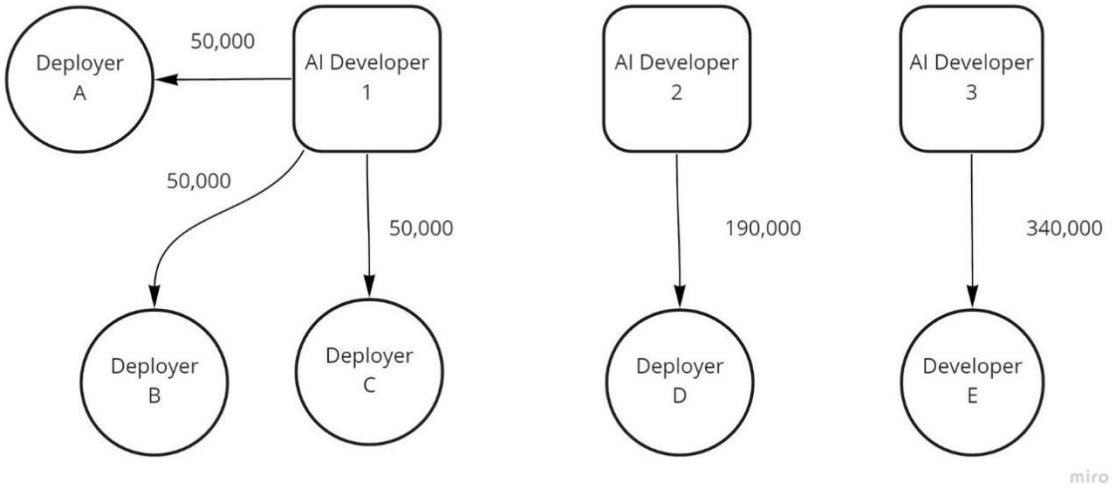
Another important assumption throughout the cost assessment is not distinguishing deployers from developers, or users from providers. The AI market is sophisticated and can be divided into five main layers, from upstream to downstream: AI and machine learning infrastructure; AI enabling technologies; horizontal applications of AI; industry-specific applications of AI; and end-users (companies and consumers) (Cognilytica, n.d.). While some large technology companies occupy the AI and machine learning infrastructure layer, numerous smaller AI developers or vendors are clustered at the layer of industry-specific applications of AI. As the regulation aims to provide similar product safety control to the end-users, it is reasonable to assume that the regulation would only be imposed on the products sold outright to end-users. Any AI developments in the upstream will finally be regulated downstream when their developments reach the market. Any development costs may also be passed onto the downstream layers, even if the development of some technologies fail in the process. This simplifying assumption allows for reliance on the AI market size to calculate the number of AI units and its projection to the future.

a. Value of an AI unit

Developing an average customised AI system may cost from USD 6,000 to USD 300,000 (or approximately EUR 5,000 to EUR 250,000)[68]. This cost estimation assumes that an AI unit costs EUR 170,000 to roughly reflect the current situation of the market and to give a better perspective for experts to evaluate these estimations. The exact reference value of an AI unit (EUR 170,000) is of secondary importance, however. Experts were asked to evaluate the study's cost estimates of the hypothetical value of average AI development cost. The essence of the evaluation is to obtain an estimate of compliance cost as a percentage of total development cost. The reference value could be another amount, but the total compliance cost estimate would be the same as long as the analysis and the industry experts respect the same reference value in estimating the costs and in evaluating these estimates.

The computation in this study relies on the assumption of a competitive market in which AI developers break even, which implies that prices just cover development costs. Consider the following economy consisting of two AI developers and four AI deployers. Developer 1 invented an AI system that could be applied by different types of companies, which then sell it to three deployers at EUR 50,000 each. Developer 2 created a customised AI system for deployer D at EUR 190,000. Developer 3 developed a very advanced customised AI system that cost EUR 340,000. Note that the average price/cost is EUR 170,000 per construction.

[68] For the cost of developing customised AI system, see: https://www.webfx.com/internet-marketing/ai-pricing.html and https://azati.ai/how-much-does-it-cost-to-utilize-machine-learning-artificial-intelligence/



The value of the market size of AI - defined as the amount of spending by customers in the market - is thus EUR 680,000. There will only be three compliance and conformity test processes, not four or five, because it is the developers who are required to send their products for examination. However, there are four AI units in the economy. Developer 3 has to pay more for its AI system, which is counted as two AI units, because it is more advanced and complex, and demands more costly compliance procedures. The essence of the calculation is not the actual values of the three AI systems but the estimated compliance costs as a percentage of a hypothetical AI unit. The percentage allows for compliance costs to be extrapolated to the whole economy. The same logic applies to the calculation of the total cost of conformity tests.

b. Other assumptions

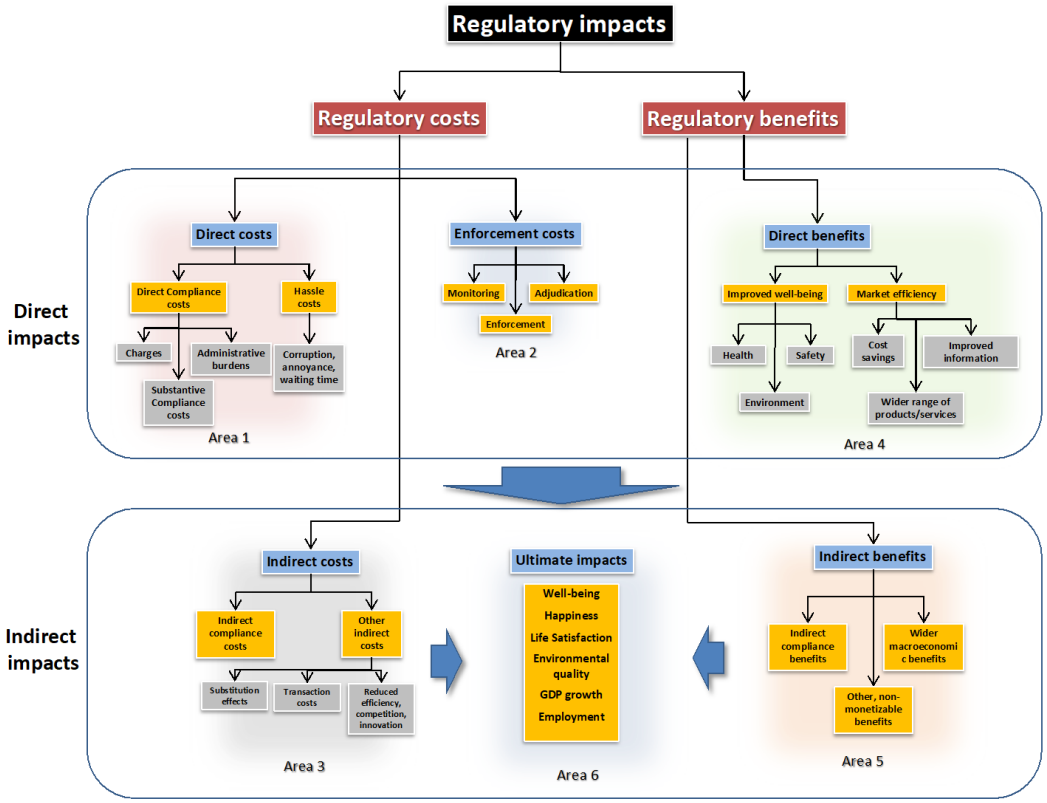
The cost assessment assigns a cost estimate to each requirement without considering specific industry characteristics. However, it is reasonable to assume scenarios where some sectors are more prepared for the proposed regulation. For example, the IT industry might have already adopted data management and governance practices that more closely resemble those required by the forthcoming regulation. The cost assessment attempts to take into account the Business-As-Usual (BAU) factor of different sectors, measured by their relative intensities of data as inputs and also outputs, by relying on estimates of data intensity per sector used in academic literature in the context of general equilibrium models, such as GTAP (van der Marel et al., 2016).

Another assumption throughout this cost estimation exercise is the linear projection of cost estimates to the future. The costs are likely to be high in the beginning phase of the inception of the proposed regulation, gradually falling to an equilibrium level over the longer term. These estimates are built on references or benchmarks in the existing market and are closer to the long-term level. Some one-off initial costs will also be discussed in the following sections.

c. Taxonomy of regulatory costs and the Standard Cost Model

Figure 19 below shows a general map of the impacts generated by legal rules, developed in Renda et al. (2013) and included in the EU Better Regulation Toolbox. As shown in the figure, regulation normally produces both direct and indirect impacts, which can, in turn, generate second-order effects (ultimate impacts).

Figure 18 - A map of regulatory costs and benefits



Source: Renda et al. (2013)

The costs assessed here refer primarily to 'Area 1', which includes 'Direct Regulatory Costs', encompassing both direct compliance costs and, as a residual category, irritation costs (or hassle costs) - these latter are typically more difficult to quantify or monetise, and are excluded here. Of the direct compliance costs, the following sub-categories are most important in this assessment:

- *Substantive compliance costs*, which encompass those investments and expenses faced by businesses and citizens in order to comply with substantive obligations or requirements contained in a legal rule. These costs can be further broken down into one-off costs (faced by regulated actors to adjust and adapt to the changed legal rule) and recurrent costs (substantive compliance costs that are borne regularly as a result of the existence of a legal rule that imposes specific periodic behaviours). These costs are calculated as a sum of capital costs, financial costs and operating costs.
- *Administrative burdens* are those costs borne by businesses, citizens, civil society organisations and public authorities as a result of administrative activities performed to comply with the information obligations (IOs) included in legal rules.

In line with the most consolidated practice, this cost assessment will follow a series of steps (Renda et al, 2013):

Step 1. *Identify the substantive duties* (SDs) generated by each of the policy alternatives, distinguishing between one-off and recurrent costs.

Step 2. *Identify information obligations* (IOs) of the five sets of potential requirements. This step will develop a conceptual map of the IOs, broken down into data requirements (DRs) and administrative activities for different stakeholder groups and sectors.

Step 3. *Estimate the population of stakeholders* that would have to comply with the potential requirements.

Step 4. *Estimate the mode of compliance* with each SD and IO by a 'normally efficient business' (including individual researchers, research organisations and institutions).

Step 5. *Estimate the BAU factor for each SD and each IO.*

Step 6. *Consider segmenting the population by creating 'case groups'* differentiated according to size (micro, small, medium, large enterprises), sector or other dimensions (level of government for public administrations, availability of internet connection for citizens, etc.). Where different case groups can be established, different notions of normal efficiency and BAU may be considered for each group (see Step 4).

Step 7. *Estimate the compliance cost* associated with each SD for each segment and each alternative, by accounting for:

- Operating and maintenance costs (OPEX), which include annual expenditures on salaries and wages, energy inputs, materials and supplies, purchased services, and maintenance of equipment. They are functionally equivalent to 'variable costs'.
- Financial costs, i.e. costs related to the financing of investment (normally considered in relation to capital costs).
- Capital Costs, 'annualised' over the period of the useful life of the equipment purchased.

Step 8. *Estimate the administrative burden* of each IO for each segment and each alternative, by accounting for:

- the *time* needed to comply with the obligation;
- the expected frequency of the IO;
- the *average salary* of the person(s) in charge of performing the underlying administrative activities;
- any *external cost* required both in terms of expert services or counselling, or acquisitions.

Step 9. *Assess whether compliance costs are likely to change* over the life of the proposed legislation. In particular, whether the impact of the costs identified is likely to change over time as a result of entry/exit of businesses, technological innovation, 'learning by doing' or any other relevant factor. This must be taken into account in a prospective analysis of regulatory costs, and – if possible – coupled with sensitivity analysis on the assumptions behind the evolution of costs over time.

Step 10. *Sum up and extrapolate all costs* to reach a total estimate. This activity will be carried out with average European data, rather than on a country-by-country basis, given the scope of the study. Importantly, two extrapolation results will be given: one related to the estimated number of new AI systems/applications introduced in

the EU market on a yearly basis; and one related to the estimated fraction of these systems/applications that could be considered 'high risk'. This will enable the Commission to work on the basis of three alternative policy options: the zero option, a 'high-risk only' option, and an alternative policy option in which the requirements are applied to all new AI systems/applications introduced in the EU market.

d. Standardised tables used in the study

The 10-step procedure described above broadly corresponds to the methodology adopted by the German government, developed with the Federal Statistical Office (Destatis). An advantage of this model is that it was accompanied by the adoption of standardised tables that allocate specific times to specific activities, differentiating each activity in terms of complexity levels. Table 18 below shows the adaptation of the table relative to businesses, which will be used for the cost assessment in this document.

Table 17 - Reference table for the assessment of compliance costs

	Time			Cost (Euros)		
	Easy	Moderate	Complex	Easy	Moderate	Complex
Administrative activities						
Familiarising oneself with the Information obligation	3	3	60	1,60	1,60	32,00
Procuring data	2	10	120	1,07	5,33	64,00
Filling in forms, labelling, classifying	3	5	30	1,60	2,67	16,00
Performing calculations	3	20	185	1,60	10,67	98,67
Checking data and inputs	1	8	60	0,53	4,27	32,00
Correcting errors	2	10	60	1,07	5,33	32,00
Processing data	3	20	240	1,60	10,67	128,00
Transmitting and publishing data	1	2	5	0,53	1,07	2,67
Internal meetings	6	60	600	3,20	32,00	320,00
External meetings	10	60	480	5,33	32,00	256,00
Payment	1	3	23	0,53	1,60	12,27
Photocopying, filing, distribution	1	2	10	0,53	1,07	5,33
Cooperating in an audit by public authorities	5	60	540	2,67	32,00	288,00
Corrections which have to be made as a result of the audit	4	30	480	2,13	16,00	256,00
Procuring additional information in case of	3	15	120	1,60	8,00	64,00
Training courses	2	30	480	1,07	16,00	256,00
Substantive costs						
Procuring goods and services						
Procuring services and/or hiring additional staff						
Supplying own services						
Adjustment of internal processes						
Supervisory measures						
Storage, inventory management, production						

Source: Authors' elaboration based on Normenkotrollrat (2018)

The translation of activities into cost estimates was obtained by using a **reference hourly wage rate of EUR 32**, which is the average value indicated by Eurostat for the Information and Communication sector (Sector J in the NACE rev 2 classification)[69].

The activities involved in complying with each of the IOs contained in a given (proposed) legal provision will be identified and colour-coded to facilitate the visual interpretation of the results. Activities of higher complexity are expected to be more costly. In particular:

- Complex activities (high-cost) are in red;
- Activities of moderate complexity (medium-cost) are in yellow;

[69] Stakeholders' feedback suggests that EUR 32 is too low, but they are operating in more advanced economies. Given the economic differences across the EU, the EU average is a reasonable reference point here.

- Easy activities (low-cost) are in green.

2. Assessing the costs of the five regulatory requirements

a. Training data

Main activities involved

This requirement, as defined in the White Paper (pp.18-19), includes the following main activities:

- Providing reasonable assurances that the subsequent use of the products or services enabled by the AI system is safe (e.g. ensuring that AI systems are trained on datasets that are sufficiently broad and cover all relevant scenarios needed to avoid dangerous situations).
- Take reasonable measures to ensure that subsequent use of AI systems does not lead to outcomes entailing prohibited discrimination, e.g. obligation to use sufficiently representative datasets, especially to ensure that all relevant dimensions of gender, ethnicity and other possible grounds of prohibited discrimination are appropriately reflected.
- Ensuring that privacy and personal data are adequately protected during the use of AI-enabled products and services. For issues falling within their respective scope, the GDPR and the Law Enforcement Directive regulate these matters.

A number of individual actions can therefore be reasonably envisaged as a result of the proposed requirement:

- Implement good practice data governance and management processes.
- Carry out a prior assessment to evaluate the availability and quality of the datasets.
- Determine whether the development of a data-driven AI system is a suitable solution to achieve the intended purpose(s) and any potential data gaps that must be addressed.
- Ensure that training datasets provide sufficiently relevant, representative, diverse, accurate, complete, timely and unbiased data for the intended purpose(s).
- Use non-personal, anonymised or synthetic datasets or, if impossible, comply with data minimisation (as per the GDPR).
- Use sufficiently broad training datasets.
- Disclose provenance and characteristics (for pre-trained datasets).
- Specify whether the system is designed to act as continuously learning after deployment, ensure that biased outputs are corrected and limitations put in place to exclude certain data from the training (for learning-based systems).
- Perform testing in a way that is proportionate to the risks and the required level of human oversight envisaged for the operation of the AI system.

Problems highlighted during the public consultation

The public consultation on the White Paper drew several related comments:

- **The requirement of training data may clash with the GDPR** (the principle of data minimisation and the right to be forgotten). GDPR rules mean that personal data cannot be collected in a lot of cases, but it may frequently be necessary in order to meet this anti-discrimination requirement.
- It puts **too much focus on past standard supervised learning from labelled data** and not enough on future AI technologies: data augmentation, transfer learning, generative adversarial methods or even model-based reinforcement learning approaches.
- It is **not feasible/possible to conduct tests for 100% of possible scenarios** nor to achieve **completely unbiased datasets**.
- **Retraining AI systems developed for a global audience with only European data would make them uneconomical** and would delay/prevent certain AI products from being made available to European consumers. It could lead to low-quality AI systems only applicable to the European market, with an obvious negative impact on consumers, innovation and business competitiveness.
- **Some systems need to be biased and are therefore trained on particular datasets.** Sometimes **biases, such as additional information/data on gender or age are intentionally created** in order to improve the learning performance in certain circumstances.
- **Assessment of training data is not the best approach to ensure the quality of the output: a more constructive approach would be to focus on testing outcomes** or to apply safeguards against biased outcomes, ensuring that outputs are within an acceptable range. This should be ex post and should not be translated into a requirement to demonstrate compliance to a regulator before launching, which would be impractical because it would require analysis and approval, creating a potential administrative backlog and significantly delaying implementation.

Identifying and measuring activities and costs

The types of activities that would be triggered by this requirement include:

- Familiarisation with the IO (one-off);
- Assessment of data availability (may require an internal meeting);
- Risk assessment (may require an internal meeting);
- Testing for various possible risks, including safety-related and fundamental rights-related risks, to then adopt and document proportionate mitigating measures;
- Anonymisation of datasets, or reliance on synthetic datasets, or implementation of data minimisation obligations;
- Collection of sufficiently broad datasets to avoid discrimination.

Table 19 summarises the likely consequences of the activities in terms of cost. These are based on the Standard Cost Model by the German Federal Government (2018). For an average process and a normally efficient firm, a reasonable cost estate for this activity is EUR 2,763.

Table 18 - Requirement 1: training data	Good practice data governance	Prior assessment of dataset quality	Suitability of data-driven solutions	Ensure quality of training dataset	Use non-personal or anonymised data	Use sufficiently broad dataset	Disclose provenance and features	Specify learning/ Correct biases	Perform proportionate testing	TOTAL
Administrative activities										
Familiarisation with IO										
Procuring data										
Filling in forms, labelling, classifying										
Performing calculations										
Checking data and inputs										
Correcting errors										
Processing data										
Transmitting and publishing data										
Internal meetings										
External meetings										
Payment										
Photocopying, filing, distribution										
Cooperating in audit by public authorities										
Corrections following audit										
Procuring additional info in case of audit										
Training courses										
Substantive costs										
Procuring goods and services	Additional data procurement to ensure sufficiently broad dataset									
Procuring services and/or hiring										
Supplying own services										
Adjustment of internal processes										
Supervisory measures										
TOTAL minutes	510	545	960	300	83	180	47,5	390	1,685	**5,180.5**
Total cost (hourly rate = EUR 32)										**€2,762.93**

b. Documents and record-keeping

Main activities involved

This requirement aims to enable the verification and enforcement of compliance with existing rules. The information to be kept relates to the programming of the algorithm, the data used to train high-risk AI systems, and, in certain cases, keeping the data themselves. The White Paper (p.19) prescribes the following actions:

- Keeping accurate records of the dataset used to train and test the AI system, including a description of the main characteristics and how the dataset was selected;
- Keeping the datasets themselves;
- Keeping documentation on programming and training methodologies, processes and techniques used to build, test and validate the AI system;
- Keeping documentation on the functioning of the validated AI system, describing its capabilities and limitations, expected accuracy/error margin, the potential 'side effects' and risks to safety and fundamental rights, the required human oversight procedures and any user information and installation instructions;
- Make the records, documentation and, where relevant, datasets available on request, in particular for testing or inspection by competent authorities.
- Ensure that confidential information is protected (e.g. trade secrets).

Problems highlighted during the public consultation

The most relevant and recurring comments received during the public consultation on this requirement include:

- Such record-keeping will complicate the development of AI by **reducing convenience and efficiency** and place a burden on companies to draw up documentation.
- Keeping vast amounts of data would be unworkable for many companies given that **AI is developed in an ongoing and iterative way.** For instance, the process of training artificial neural networks is a complex process that requires evaluation of many different model parameters and use of different data and different software versions, which would make control and recording using conventional methods difficult.
- It will be very complex and costly **for already applied AI** systems, as numerous datasets cannot be recreated.
- If keeping of data could potentially reveal details of AI systems and underlying code, this could **undermine privacy and trade secrets, infringe on intellectual property rights, and heighten cybersecurity risks, privacy and data manipulation risks**.
- It raises potential problems and conflicts with **copyright law** (e.g. copyrighted datasets authorised for only short-term access) and also with the GDPR, in particular with the right to be forgotten and with privacy rights. It also conflicts with the targets of the European Green Deal as it would consume significant storage resources (**environmental cost**).
- Other specific AI system learning techniques are built to protect privacy (federated learning) and **disclosure obligations** could undermine this crucial goal.

- Edge computing is not considered - this would destroy the **privacy benefits of on-device processing** because it would effectively force data to be collected and stored centrally. This mandate would also prevent utilising off-the-shelf and open-source models.
- There are no common data naming conventions, no formatting standards or concurrent versioning systems used for data. This would make regulation in this area challenging, given the **vast datasets used in AI development and the lack of an established standard** to allow these datasets to be shared or reviewed in a way that would be meaningful for an assessment.

Identifying and measuring activities and costs

The first activity required (keeping records on the training data) would overlap with activities already foreseen under Requirement 1 (see above) and is thus excluded here to avoid double counting. The remaining obligations were broken down into administrative activities and data requirement, as shown in the table below. Feedback from stakeholders stresses the need for a dedicated data officer managing data and records and ensuring compliance, although the cost could be shared among different products. The data officer must be well-trained, with the necessary legal knowledge. For an average process and a normally efficient firm, a reasonable cost estimate per AI product for this activity is **EUR 1,190,** together with the cost of 0.05 full-time equivalent (FTE) data officer, at EUR 3,200 per year.

Table 19 - Requirement 2 documents and record-keeping	Records on the dataset used to train and test the system	Keep datasets	Documents on programming and training, processes and techniques	Keeping documentation on the functioning of the validated AI system	Make records and data available upon request	Protect confidential information	TOTAL
Familiarisation with IO	Overlap with Requirement 1						
Procuring data							
Filling in forms, labelling, classifying							
Performing calculations							
Checking data and inputs							
Correcting errors							
Processing data							
Transmitting and publishing data							
Internal meetings							
External meetings							
Payment							
Photocopying, filing, distribution							
Cooperating in audit by public authorities (freq = 0.5)							
Corrections following audit							
Procuring additional information in case of audit (freq = 0.5)							
Training courses							
Procuring goods and services	Legal advice?						
Procuring services and/or hiring additional staff	0.05 FTE data officer – EUR 3,200/year						
Total minutes	0	23	215	875	338	780	2,231
Total cost (hourly rate = EUR 32)							EUR 1,190
Total cost							EUR 4,390

c. Provision of information

Main activities

Beyond the record-keeping requirements discussed earlier, adequate information is required on the use of high-risk AI systems. According to the White Paper (p.20), the following requirements could be considered:

- Ensuring clear information is provided on the AI system's capabilities and limitations, in particular the purpose for which the system is intended, the conditions under which it can be expected to function as intended, and the expected level of accuracy in achieving the specified purpose. This information is especially important for deployers of the systems, but it may also be relevant to competent authorities and affected parties.
- Citizens should be clearly informed when they are interacting with an AI system and not a human being.

In practice, the future regulation could include several required actions:

- **Ensuring clear information on the AI system's capabilities and limitations**, the purpose for which the system is intended, the conditions under which it can be expected to function as intended, and the expected level of accuracy in achieving the specified purpose. This could include information on:
    - Identity and contact details of the provider;
    - Purpose and key assumptions/inputs to the system;
    - What the model is designed to optimise for, and the weight accorded to the different parameters;
    - System capabilities and limitations;
    - Context and the conditions under which the AI system can be expected to function as intended and the expected level of accuracy/margin of error, fairness, robustness and safety in achieving the intended purpose(s);
    - Potential 'side effects', the safety and fundamental rights risks posed by the AI system and any known and foreseeable circumstances that may impact on the accuracy, fairness, robustness and safety of the system; specific conditions and instructions on how to operate the AI system, including information about the required level of human oversight (if any) and any other mitigating and precautionary measures that users shall take to avoid or minimise the safety and fundamental rights risks;
    - For users: concise, clear, non-technical and intelligible information specifying the identity and the contact details of the user and, where applicable, of their authorised representative;
    - For users: information on whether an AI system is used for interaction with humans (unless immediately apparent from the context or the AI system is integrated as optimisation techniques);
    - For users: information on whether the system is used as part of a decision-making process that significantly affects the person;
    - For users: at the request of an affected person (unless required in all circumstances by law), an explanation of the individual decision;

  - For users: at the request of an affected person, immutable auditable logs of how the AI system has performed in the particular case of the complainant, and available remedies under applicable law;
  - For users: a summary of the DPIA carried out following Article 35 of the GDPR.
- **Inform citizens** when they are interacting with an AI system and not a human being (apart from situations where it is immediately obvious to citizens that they are interacting with AI systems).
- **Design AI systems** in a transparent and explainable way to enable human operators to understand and control how the AI system achieves the output and to be able to explain that output to affected persons, notified bodies or competent supervisory authorities.

Problems highlighted during the public consultation

According to some of the stakeholders responding to the public consultation, the information required is already sufficiently provided in most cases, especially in B2B relations. There should therefore be a differentiation between transparency requirements for AI applications being used in consumer-facing vs. B2B products and services, where there is no reason to share such information, except where it is deemed to be critical for public interest. Excessive sharing obligations might put intellectual property rights at risk, or indeed contractual arrangements between business partners.

Identifying and measuring activities and costs

The types of activities that would be triggered by this requirement include:

- Provide information on the AI system's characteristics, such as
  - Identity and contact details of the provider;
  - Purpose and key assumptions/inputs to the system;
  - What the model is designed to optimise for, and the weight accorded to the different parameters;
  - System capabilities and limitations;
  - Context and the conditions under which the AI system can be expected to function as intended and the expected level of accuracy/margin of error, fairness, robustness and safety in achieving the intended purpose(s);
  - Potential 'side effects' and safety/fundamental rights risks;
  - Specific conditions and instructions on how to operate the AI system, including information about the required level of human oversight.
- **Provide information** on whether an AI system is used for interaction with humans (unless immediately apparent).
- **Provide information** on whether the system is used as part of a decision-making process that significantly affects the person.
- **Design AI systems** in a transparent and explainable way.
- **Respond to information queries** to ensure sufficient post-purchase customer care. This activity was stressed by stakeholders with experience of GDPR compliance.

Table 21 summarises the likely consequences of these activities in terms of cost. Given the overlaps with activities foreseen under other requirements, only the familiarisation with the specific IOs and their compliance are included, rather than the cost of the underlying activities. Nevertheless, this requirement may also entail changes in the design of the AI system to enable explainability and transparency, an aspect about which there is limited academic literature, and that can only be verified with market players.

For an average process and a normally efficient firm, a reasonable cost estate for this activity is **EUR 3,627**.

Table 20 - Requirement 3: information provision	Identity/ contact of the provider	Purpose and key assumptions	Model optimisation and parameters	Capabilities and limitations	Context and conditions of use, accuracy and fairness	Potential effects on safety and fundamental rights risks	Instructions of use, including required oversight	Info on system interaction with humans	Info on system impact on persons	Info queries	
Administrative activities											
Familiarisation with IO											
Procuring data											
Filling in forms, labelling, classifying											
Performing calculations											
Checking data and inputs											
Correcting errors											
Processing data											
Transmitting and publishing data											
Internal meetings											
External meetings											
Payment											
Photocopying, filing, distribution											
Cooperating in audit by public authorities											
Corrections following audit											
Procuring additional information in case of audit											
Training courses											
Substantive costs											
Procuring goods and services	Legal advice on safety and fundamental rights										
Procuring services and/or hiring additional staff											
Adjustment of internal processes	Changes in system design to enable explainability and transparency										
TOTAL minutes	7	20	95	78	635	1,505	1,500	615	545	**1,800**	**6,800**
Total admin cost (hourly rate = EUR 32)											**EUR 3,627**

d. Human oversight

Main activities involved

The White Paper acknowledges that the type and degree of human oversight may vary from one AI system to another (European Commission, 2020a, p.21). It will depend, in particular, on the intended use of the AI system and the effects of that use on affected citizens and legal entities. For instance:

- Output of the AI system does not become effective unless it has been previously reviewed and validated by a human (e.g. the rejection of an application for social security benefits may be taken by a human only).
- Output of the AI system becomes immediately effective, but human intervention is ensured afterwards (e.g. the rejection of an application for a credit card may be processed by an AI system, but human review must be possible afterwards).
- Monitoring of the AI system while in operation and the ability to intervene in real time and deactivate (e.g. a stop button or procedure is available in a driverless car when a human determines that car operation is not safe).
- In the design phase, by imposing operational constraints on the AI system (e.g. a driverless car shall stop operating in certain conditions of low visibility when sensors may become less reliable, or shall maintain a certain distance from the vehicle ahead in any given condition).

This makes it rather difficult to associate a cost measure with a specific type of conduct, as the latter may change significantly depending on the case. Generally, however, this requirement may entail the following procedures:

Adopting **technical and organisational measures** tailored to the intended use of the AI system, to be assessed after the design phase of an AI system, right through the point at which the system is released to the market. These may include:

- **Measures to prevent and mitigate automation bias**, in particular for AI systems used to assist humans;
- Measures to **detect and safely interrupt anomalies**, **dysfunctions, unexpected behaviour**.

Problems highlighted during the public consultation

- Some respondents noted that human oversight can be especially detrimental in cases that require (or benefit from) very fast response times (e.g. avoiding an accident or high-frequency trading) and in cases where capabilities can be made accessible at a much cheaper cost than with continuous human involvement. It could deter the development and introduction of fully automated technologies in Europe, potentially leading to delays.
- This requirement should not counteract the advantages gained by using AI systems: in some cases, the accuracy of outputs could even be undermined by human interventions.

Identifying and measuring activities and costs

This requirement is not easily applied to standardised tables, due to the scale of uncertainty about the scope, measures and type of oversight involved. Possible activities involved in complying with this requirement can be drawn from the ALTAI questions (AI HLEG, 2020):

- **Monitoring** the operation of the AI system, including detection of anomalies, dysfunctions, and unexpected behaviour;
- Ensuring **timely human intervention**, such as a 'stop' button or procedure to safely interrupt the operation of the AI system;
- Conducting **revisions in the design** and functioning of the currently deployed AI, including measures to prevent and mitigate automation bias;
- Overseeing the **overall activity of the AI system** (including its broader economic, societal, legal and ethical impacts);
- Implementing **additional hardware/software/systems** assisting staff in the abovementioned tasks to ensure meaningful human oversight over the entire AI system lifecycle;
- Implementing **additional hardware/software/systems** to meaningfully explain to users that decisions, content, advice or outcomes is the result of an algorithmic decision, and to prevent end users over-reliance on the AI system.

Estimating the average cost of the human oversight requirement is substantially complicated by the number of assumptions that are needed. Among the key unknowns are:

- How many currently operating AI systems have insufficient human oversight?
- What kind of human oversight will be considered meaningful depending on the circumstances and the type of use case (HITL, HOTL, HIC)?
- Whether compliance oversight requires a redesign of the AI system itself.

It is therefore assumed that the following actions would be needed to comply with the human oversight requirement:

- Hiring dedicated staff: e.g. 0.1 FTE experienced data scientist;
- Implementing software upgrades, including AI (e.g. for anomaly detection);
- Providing extensive staff training.

It is further assumed that action will be needed for all AI systems. However, in computing the BAU factor, the estimated amount was discounted in order to account for existing practices in the market.

The total estimate was **EUR 7,764.**

Table 21 - Requirement 4: Human oversight	Hiring dedicated staff	Software upgrades	Staff training	TOTAL
Administrative activities				
Familiarisation with IO				
Procuring data				
Filling in forms, labelling, classifying				
Performing calculations				
Checking data and inputs				
Correcting errors				
Processing data				
Transmitting and publishing data				
Internal meetings				
External meetings				
Payment				
Photocopying, filing, distribution				
Cooperating with audit by public authorities				
Corrections following audit				
Procuring additional information in case of audit				
Training courses				
Procuring goods and services	Purchase of additional software EUR 500			
Procuring services and/or hiring additional staff	0.1 FTE data scientist – EUR 6,400/year			
Supplying own services				
Adjustment of internal processes				
Supervisory measures				
Storage, inventory management, production				
TOTAL minutes	600	540	480	**1,620**
Total admin cost (hourly rate = EUR 32)				**EUR 864.00**
Total cost				**EUR 7,764.00**

e. Robustness and accuracy

Main activities involved

According to the White Paper on Artificial Intelligence (European Commission, 2020a, p. 20), 'AI systems must be technically robust and accurate if they are to be trustworthy. Such systems, therefore, need to be developed in a responsible manner and with ex ante due and proper consideration of the risks they may generate. Their development and functioning must be such to ensure that AI systems behave reliably as intended. All reasonable measures should be taken to minimise the risk of harm.' Accordingly, the following elements could be considered:

- Requirements ensuring that the AI systems are robust and accurate, or at least correctly reflect their level of accuracy, during all lifecycle phases;
- Requirements ensuring that outcomes are reproducible;
- Requirements ensuring that AI systems can adequately deal with errors or inconsistencies during all lifecycle phases;
- Requirements ensuring that AI systems are resilient against overt attacks and against more subtle attempts to manipulate data or algorithms, and that mitigating measures are taken in such cases.

Problems highlighted during the public consultation

- Due to the particularly large amount of data required to train AI algorithms, assessing the accuracy and quality of those algorithms is essentially an impossible task and may **reduce the effectiveness** of the whole AI-based system, both in speed and quality.
- The quality requirements for different deployments **vary significantly between domains**.
- It should be accepted and understood that AI will make mistakes and 100% accuracy is not possible. It should be evaluated **relative to human accuracy rates**: there should not be a higher standard set for AI than for human decision-making.
- The reproducibility requirement is not always appropriate and in the interest of the user - when new versions of AI systems come at short intervals, the requirement ensuring that outcomes are reproducible imply that all intermediate versions must be kept available. It is often **not possible to achieve reproducibility -** for AI systems that change over time, it would require reproducing the entire dynamic environment and the entirety of data used to train the model. In practice, this could lead to AI systems being built on only very basic techniques, as reproducibility of more complex systems would not be possible.

Identifying and measuring activities and costs

Compliance with this requirement will likely entail technical and organisational measures tailored to the intended use of the AI system, to be assessed from the design phase of an AI system, right through to the point at which the system is released to the market. It includes measures to prevent and mitigate automation bias, in particular for AI systems used to provide assistance to humans, as well as measures to detect and (safely) interrupt anomalies, dysfunctions and unexpected behaviour.

There are two types of general requirements, one related to accuracy and another to robustness. For every single AI product, the following activities will apply.

On accuracy:

- Familiarisation with accuracy requirements;
- Calculating an established accuracy metric for the task;
- Writing an explanation of the accuracy metric that can be understood by laypeople;
- Procure external test datasets and calculate additional required metrics.

On robustness:

- Familiarisation with robustness requirements;
- Brainstorming possible internal limitations and external threats of the AI model;
- Describing limitations of the AI system based on knowledge of the training data and algorithm;
- Conducting internal tests against adversarial examples (entails possible retraining, changes to the algorithm, 'robust learning');
- Conducting internal tests against model flaws (entails possible retraining, changes to the algorithm);
- Conducting tests with external experts (e.g. workshops, audits);
- Conducting robustness, safety tests in real-world conditions (controlled studies, etc.).

Additional labour is likely to be necessary to ensure that development complies with requirements and to keep records of testing results for future conformity assessment.

Table 22 - Requirement 5: Robustness and Accuracy	Accuracy	Robustness	Security	TOTAL
Familiarisation with IO				
Procuring data				
Filling in forms, labelling, classifying				
Performing calculations				
Checking data and inputs				
Correcting errors				
Processing data				
Transmitting and publishing data				
Internal meetings				
External meetings				
Payment				
Photocopying, filing, distribution				
Cooperating with audit by public authorities				
Corrections following audit				
Procuring additional information in case of audit				
Training courses				

Procuring goods and services	Pen-testing costs approx. EUR 5,000-10,000; External security services (e.g. Red Team) cost around EUR 200/hr (est. 30 hours) = EUR 5,000			
Procuring services and/or hiring additional staff	0.05 FTE data scientist – EUR 3,200/year			
Supplying own services				
Adjustment of internal processes	Possible redesign of the business model to ensure reproducibility? Possible additional cost of data/information storage?			
Supervisory measures				
Storage, inventory management, production				
TOTAL minutes	1,205	2,405	1,140	4,750
Total admin cost (hourly rate = EUR 32)				EUR 2,533.33
Total cost				EUR 10,733.33

3. Total compliance cost of the five requirements for each AI product

Table 24 summarises the main activities to be performed in order to comply with the five requirements and the associated costs for each AI unit. These cost estimates still include the BAU factor, which will vary depending on the extent to which the activities are already performed by the regulated entities as part of their internal practice, adherence to industry standards, or existing legislation (for estimates where the BAU factor is calculated and subtracted). The estimated annual labour compliance cost for a single AI product is EUR 10,977. Together with the purchase of external data and services, as well as hiring additional staff, this cost may rise to **EUR 29,277**. The annual compliance cost is **17.22%** of the value of a reference AI unit (EUR 170,000), which is a reasonable amount (see comparative compliance costs in the corresponding section).

These estimates are meant to be representative of the activities that would be needed for an average firm to comply with the five regulatory requirements along the lifecycle of an average AI application. As such, they give an approximation of the costs incurred by compliance with the regulatory requirements. In order to ensure that the activities, times and costs estimated are reasonable, a dedicated workshop was held with a group of stakeholders from different sectors, including bot developers and deployers of AI systems and large and small businesses. Information was also collected through a questionnaire that specifically aimed at validating the parameters and assumptions underlying these estimates.

Table 23 -Cost of all five requirements	Training data	Documents and record-keeping	Information provision	Human oversight	Robustness and accuracy	TOTAL
Administrative activities						
Familiarisation with IO	9	5	10	1	3	28
Procuring data	3		6		1	10
Filling in forms, labelling, classifying	2	2	3			7
Performing calculations	2				2	4
Checking data and inputs	6	2	1		2	11
Correcting errors	3	1			2	6

Processing data	5	3	3		2	11
Transmitting and publishing data		3	9			12
Internal meetings	5		6	1	3	15
External meetings		1	3	1	1	6
Payment						
Photocopying, filing, distribution						
Cooperating with audit by public authorities		1				1
Corrections following audit						
Procuring additional information in case of audit		1				1
Training courses	1	1	1	1	2	6
Additional costs						
Procuring goods and services	Purchasing additional data (EUR 500) + additional legal advice					
Procuring services and/or hiring additional staff						
Supervisory measures	Security testing services (EUR 5,000) 0.2 FTE staff– EUR 12,800/year					
Total minutes	5,180.5	2,231	6,800	1,620	4,750	20,581.5
Total admin cost (hourly rate = EUR 32)						EUR10,976.8
Total cost						EUR 29,276.8

4. Projection to the population

As discussed above, the cost estimates of each requirement of the regulation are based on an average hypothetical AI system. To recap, this cost assessment takes USD 200,000 or **EUR 170,000** as the reference value of an AI.

By dividing total AI market size by the reference AI value, a value of units of AI employed in the market is obtained, which is then multiplied by the compliance cost per unit of reference AI.

The study team used a series of available estimates on the size and evolution of the AI market globally[70]. Analysts may use different definitions of AI, making reported amounts difficult to compare directly, but they are nevertheless a useful guide. The European share of the global AI market was assumed to be 22%, based on its share in the AI software market (Statista, 2019).

Forecasts made after the COVID-19 pandemic are significantly higher, enabling the use of two types of forecasts - those published before February 2020 (pre-Feb 2020) and those after (post-Feb 2020). One of each estimate will be used to give a lower and higher bound, respectively (see detailed explanation in the Annex). The 'high growth' scenario is believed to be more likely, given the agreement between some of the most recent estimates. They have accounted for the latest developments, such as a push to digitisation due to the imposed movement restrictions. The lower bound is used as a precaution against the event of a 'digital bubble'.[71]

[70] ReportLinker, OECD (based om Crunchbase), CB Insights, McKinsey Global Institute, International Data Corporation, Grand View Research, Allied Market Research, Statista/Tractica, OMDIA/Tractica, UBS, Markets and Markets, McKinsey, International Data Corporation, and Zion Market Research.

[71] i.e. the overestimation of the the speed of the economy's adoption of digital tools because of temporary flunctuations in digital uptake resulting from the coronavirus pandemic-related movement redtrictions

To avoid combining heterogeneous estimates, two estimates of global AI market size were selected, from Allied Market Research (2018) and Grand View Research (2020). The rate of exponential growth was calculated from the initial and final values for the forecast period. With these in mind, the average compound annual growth rate (CAGR)[72] was deduced and the years in between were estimated. In line with Tractica/Statista, this report assumes that the EU share of the global investment is 22%.

Table 25 reports the AI investment values from 2020 to 2025 in EUR million, assuming an exchange rate of USD 1 = EUR 0.85. These two sources of information will thus form the foundation of the upper (Grand View Research) and lower (Allied Marker Research) compliance cost estimates of this report[73]. As the Grand View Research report was more recently published, it was expected to be more accurate. The lower-bound estimate is useful as a more conservative reference.

[72] CAGR=(Market value in the final time periodMarket value in the initial time period)1/(number of time periods)

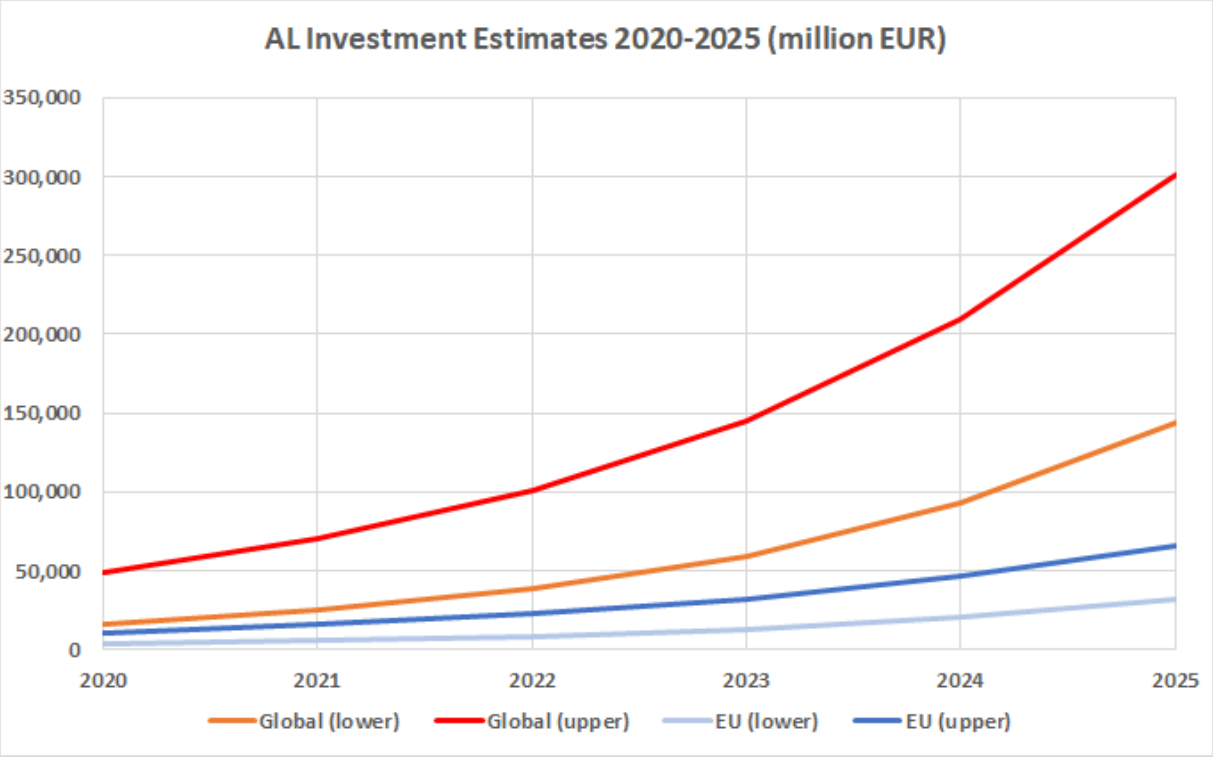
[73] The study team obtained two point estimates of each research report and interpolated the values in between by assuming exponential growth.

Table 24 - AI investment estimates (EUR million), 2020-2025

AI investments	2020	2021	2022	2023	2024	2025
Global (Grand View)	48,804	70,231	101,064	145,433	209,283	301,163
EU (Grand View)	10,737	15,451	22,234	31,995	46,042	66,256
Global (Allied Market)	15,788	24,566	38,224	59,476	92,545	144,000
EU (Allied Market)	3,473	5,404	8,409	13,085	20,360	31,680

Source: Authors' extrapolation from Allied Market Research, Grand View Research and Tractica (to estimate EU share)

Figure 19 - Global and EU AI investment, 2020-2025



Source: Authors' computation based on Allied Market Research, Grand View Research and Tractica.

a. Total compliance costs (no BAU considered)

Multiplying the estimated number of AI units by EUR 29,276.8 (estimated annual compliance cost, see above) obtains a projection of the total compliance costs for the EU economy and the global economy from 2020 to 2025. Table 26 summarises the cost estimates. By 2025, without considering the BAU factor, total compliance costs may range from EUR 5.5 billion to EUR 11.4 billion in the EU, and EUR 24.8 billion to EUR 51.9 billion globally.

Table 25 - Projection to the population 2020-2025 (EUR million)

100% coverage	2020	2021	2022	2023	2024	2025
Global (high)	8,404.93	12,094.98	17,404.91	25,046.01	36,041.94	51,865.17
Global (low)	2,718.94	4,230.64	6,582.74	10,242.78	15,937.85	24,799.21
EU (high)	1,849.08	2,660.90	3,829.08	5,510.12	7,929.23	11,410.34
EU (low)	598.17	930.74	1,448.20	2,253.41	3,506.33	5,455.83

b. EU compliance costs by sector

A projection to different sectors is more challenging. From a report by Ipsos (2020) for the European Commission[74], estimates were obtained on the percentage of firms using at least one AI system, across 17 sectors. Additionally, firms that do not currently use AI indicate whether they plan to do so in the next two years (Ipsos, 2020). Assuming that the growth beyond 2022 would slow, the adoption rate over the next four years (2023-2026) was halved. The total compliance costs were then broken into sectors. Multiplying the AI adoption rate by each sector's gross value added[75] yielded the gross value added of firms using at least one AI system. The value as a proportion of the sum of values is then taken as the weight for each sector in the subsequent cost allocation. The same process applies to each year from 2023 to 2025. Table 27 relies on the lower bound of the cost estimates. No comparable data were found for the global market, thus the cost allocation was computed for the EU only.

Table 26 - Weights to divide total compliance cost

Year	2020	2021	2022	2023	2024	2025
Accommodation, food	0.059	0.058	0.057	0.057	0.057	0.057
Agriculture, forestry and fishing	0.019	0.019	0.019	0.019	0.020	0.020
Construction	0.053	0.054	0.055	0.055	0.055	0.055
Education	0.069	0.069	0.070	0.070	0.070	0.070
Finance, insurance	0.056	0.062	0.066	0.068	0.070	0.071
Human health	0.066	0.066	0.066	0.066	0.066	0.066
IT	0.076	0.069	0.065	0.063	0.061	0.059
Manufacturing	0.210	0.204	0.200	0.199	0.197	0.196
Oil and gas	0.004	0.004	0.004	0.004	0.003	0.003
Other technical/scientific sectors	0.080	0.080	0.081	0.081	0.081	0.081
Real estate	0.131	0.132	0.133	0.134	0.134	0.134
Recreation activities	0.014	0.014	0.013	0.013	0.013	0.013
Social work	0.030	0.028	0.026	0.025	0.025	0.024
Trade, retail	0.047	0.050	0.051	0.052	0.053	0.053
Transport	0.049	0.052	0.054	0.055	0.056	0.056
Waste management	0.006	0.007	0.008	0.009	0.009	0.009
Water and electricity supply	0.032	0.031	0.031	0.031	0.031	0.031

One important assumption behind the calculation is the equivalence of treatment between developers and deployers. While it is true that the IT industry has 'paid' a higher amount for the compliance costs, the study assumes that these costs would be passed downstream and finally shared equally between developers, deployers and end users. Therefore, costs

[74] The study is based on 8,661 interviews within the EU-27.

[75] Data for GVA by sector from Eurostat [nama_10_a64], 2017.

per sector will not be assessed strictly on their use of AI, but on a combination of their AI adoption rate and their gross value added.

BAU factors

The previous cost estimation does not take into account various levels of preparedness for the forthcoming AI regulation. As stated, a unit of AI costs EUR 29,277 and the cost is distributed **evenly** among different sectors, according to their AI adoption rate and their size of gross value added. The even distribution assumption could be challenged on the basis that some sectors are already compliant with other regulations in the digital single market, such as the GDPR. Overlapping activities between existing regulations and the new AI regulation is the BAU factor. To integrate the BAU factor into these cost estimates, the study team used the data intensity index of van der Marel et al. (2016), which assumes that the higher the data intensity, the better the preparedness. Since 2016 - and the subsequent implementation of the GDPR in May 2018 - industries have reasonably strengthened their data protection and storage capacity. Yet, the new requirements would still impose additional costs, given the additional requirements beyond those in the GDPR. A simple approximation of the preparedness of each sector ranks all sectors according to their data intensities, fixes the maximum amount of costs that could be avoided, and computes the costs avoided by each sector accordingly. For example, the IT industry's data intensity is 0.318, while the transport industry is 0.032. Logic suggests that the more data-intensive the sector, the greater its familiarity with managing and restoring data, and awareness of related regulations. Although sectors may invest in AI without being a developer, discussions with stakeholders suggest that differences in preparedness exist between sectors because any customised AI systems (as collaborations between upstream developers and downstream retail-level companies) will involve the provision of data from both sides. It is thus reasonable to assume AI investment by downstream retail-level companies to bear the cost of relatively insufficient preparedness of data-related regulations.

The data intensity index was computed for the year 2016 and was therefore adjusted by the Digital Economy and Society Index (DESI)[76]. More specifically, the index of the Integration of Digital Technology (Dimension 4) was taken as the adjustment factor. The EU has gained, on average, 24.77% in the dimension of Integration of Digital Technology (rising from 33.1 to 41.3). The DESI does not provide information by sector and it would be incorrect to assume that all sectors' digital technology adoption or data intensity has grown by 24.77% evenly over the same period, as some sectors adopted digital technology earlier than others. Thus, sectors are first classified into four categories according to their data intensity index in 2016, with each category then assumed to grow differently in terms of digital technology, to allow for catching up of those lagging, while keeping the average growth rate equal to 24.77% and retaining the ranking of data intensity of 2016. Although exact data are unavailable, increases in the use and trading of data in different sectors are well-documented (Spiekermann, 2019). Several sectors are particularly involved in the digital transformation, such as banking and insurance, media, healthcare, education and manufacturing (Maruti Techlabs, 2017). Table 28 details the estimates of data intensities in 2020.

Table 27 - Data intensity estimates for 2016 and 2020

Sector	Data intensity 2016	Growth (%)	Data intensity 2020
IT	0.318	16.25	0.369675

[76] https://digital-agenda-data.eu/charts/desi-components#chart={%22indicator%22:%22desi_4_idt%22,%22breakdown-group%22:%22desi_4_idt%22,%22unit-measure%22:%22egov_score%22,%22time-period%22:%222020%22}

Other technical /scientific sectors	0.069	24	0.08556
Finance, insurance	0.05	24	0.062
Accomodation, food	0.048	24	0.05952
Recreation activities	0.048	24	0.05952
Education	0.04	30	0.052
Human health	0.04	30	0.052
Social work	0.04	30	0.052
Trade, retail	0.037	30	0.0481
Water and electricity supply	0.034	30	0.0442
Waste management	0.034	30	0.0442
Transport	0.032	30	0.0416
Manufacturing	0.024	30	0.0312
Construction	0.024	30	0.0312
Real estate	0.024	30	0.0312
Oil and gas	0.011	40	0.0154
Agriculture, forestry and fishing	0.007	40	0.0098

Denote the ratio between the two data intensities by preparedness score. Assuming that at most **50%** of the compliance costs could be reduced through the BAU factor, the IT industry and the transport industry are expected to pay 50% less and 5.63% less compliance cost respectively[77]. Over time, companies are expected to learn from experience and become familiar with their obligations. For example, referring to Table 28, the costs of familiarisation with IOs, data procurement and correction of errors fall over time. The preparedness levels of sectors are assumed to increase over time, starting from 2020 (and respecting the existing levels of preparedness of different sectors while allowing sectors to catch up with the IT sector). The yearly increase in sectors' preparedness levels is assumed at 5% in 2021, 10% in 2022, 50% in 2023, 75% in 2024 and 100% in 2025[78]. It is reasonable to assume that preparedness across sectors will gradually increase and catch up with the IT industry, as it takes time for companies in different sectors to familiarise themselves with the new requirements. Table 29 shows the estimates of preparedness in different sectors from 2020 to 2025.

Table 28 - Estimated preparedness across sectors

Preparedness level	2020	2021	2022	2023	2024	2025
IT	1	1	1	1	1	1
Other technical /scientific sectors	0.231447	0.243019	0.267321	0.400981	0.701717	1
Finance, insurance	0.167715	0.176101	0.193711	0.290566	0.508491	1
Accomodation, food	0.161006	0.169057	0.185962	0.278943	0.488151	0.976302

[77] Preparedness of the transport industry = 0.0416/0.369675 = 0.1125. Saving of the transport industry = (50% X 0.1125) = 5.63%.

[78] Given the reasonable assumption that the AI regulation would probably be enacted in 2022 and thus the biggest increase should occur in 2023.

Recreation activities	0.161006	0.169057	0.185962	0.278943	0.488151	0.976302
Education	0.140664	0.147697	0.162467	0.243701	0.426476	0.852952
Human health	0.140664	0.147697	0.162467	0.243701	0.426476	0.852952
Social work	0.140664	0.147697	0.162467	0.243701	0.426476	0.852952
Trade, retail	0.130114	0.13662	0.150282	0.225423	0.39449	0.788981
Water and electricity supply	0.119564	0.125543	0.138097	0.207145	0.362505	0.725009
Waste management	0.119564	0.125543	0.138097	0.207145	0.362505	0.725009
Transport	0.112531	0.118158	0.129974	0.19496	0.341181	0.682362
Manufacturing	0.084398	0.088618	0.09748	0.14622	0.255886	0.511771
Construction	0.084398	0.088618	0.09748	0.14622	0.255886	0.511771
Real estate	0.084398	0.088618	0.09748	0.14622	0.255886	0.511771
Oil and gas	0.041658	0.043741	0.048115	0.072173	0.126302	0.252605
Agriculture, forestry and fishing	0.02651	0.027835	0.030619	0.045928	0.080374	0.160749

Table 30 presents the compliance costs by sector based on the upper-bound (most recent) estimates for AI market size. In general, if the BAU factor does not vary over time, the yearly BAU discount rate is 9%. Applying the gradual rise formula again, the BAU discount goes up to **36.37%** in 2025.

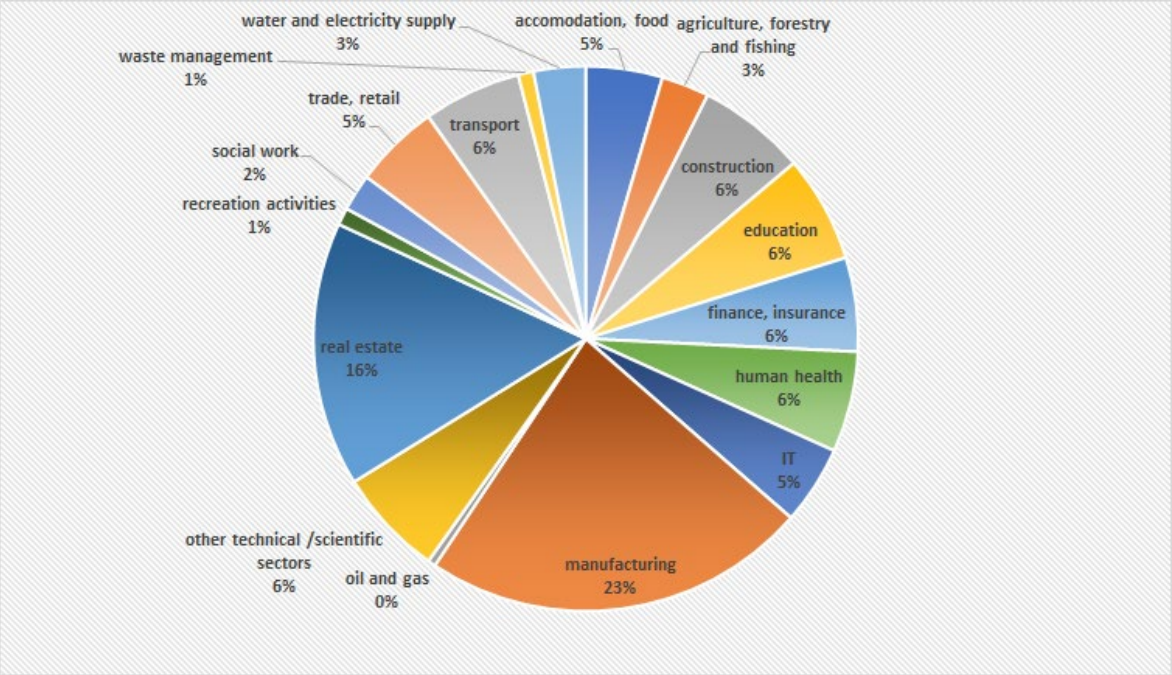
Table 29 - Compliance costs by sector (max. BAU factor = 50%)

Sector	2020	2021	2022	2023	2024	2025
Accommodation, food	101	142	199	271	340	330
Agriculture, forestry and fishing	34	50	73	105	149	206
Construction	94	137	199	280	381	469
Education	118	171	246	339	438	460
Finance, insurance	94	150	229	321	412	405
Human health	114	163	233	321	413	434
IT	70	92	124	173	242	339
Manufacturing	371	519	729	1015	1364	1665
Oil and gas	8	10	14	19	26	33
Other technical /scientific sectors	131	188	268	357	418	464
Real estate	232	336	485	683	927	1140
Recreation activities	24	33	46	62	77	74
Social work	52	68	92	123	155	159
Trade, retail	82	123	182	255	336	368
Transport	85	129	193	272	365	422
Waste management	11	18	29	42	57	66
Water and electricity supply	55	78	111	153	200	224
Total (EUR million)	**1,674**	**2,409**	**3,451**	**4,788**	**6,299**	**7,261**

Source: Authors' computation

Figure 21 shows the percentage cost burdens of the 17 sectors in 2025. The most affected industry is manufacturing, followed by real estate, finance and insurance, education and health.

Figure 20 - Compliance cost distribution in 2025 (EU)



Total compliance costs (BAU considered)

In the absence of any data on AI investment distribution across sectors for the global economy, the BAU factor discount rates are simply applied to the corresponding years, which are computed using the EU data. Table 31 reports the projection up to 2025 for the EU and the global economy. Compared to no BAU factor taken into account, the total compliance costs are **36.37%** lower in 2025. In other words, the compliance costs of an AI unit (**EUR** 170,000) would fall to EUR 18,629, or 11% of the reference value.

Table 30 - Projection to the population, 2020-2025, BAU considered (EUR million)

100% coverage	2020	2021	2022	2023	2024	2025
Global (high)	7,610.92	10,951.49	15,688.55	21,762.20	28,633.27	33,003.17
Global (low)	2,462.08	3,830.67	5,933.60	8,899.84	12,661.71	15,780.39
EU (high)	1,674.40	2,409.33	3,451.48	4,787.68	6,299.32	7,260.70
EU (low)	541.66	842.75	1,305.39	1,957.96	2,785.58	3,471.69

Source: Authors' own computation

c. High-risk only regulation

The previous cost calculation does not differentiate high-risk AI from low-risk AI - the actual cost will be much lower if only certain types of AI are regulated. The study assumes that **10%** of all AI systems are high-risk and their unit price is the same. Table 32 and Figure 22 summarise the compliance costs for the EU and global economies of an AI regulation that covers only 10% of AI investment/units, with the BAU factor taken into account. In 2022, when the proposed regulation is assumed to become effective, the private sector of the EU

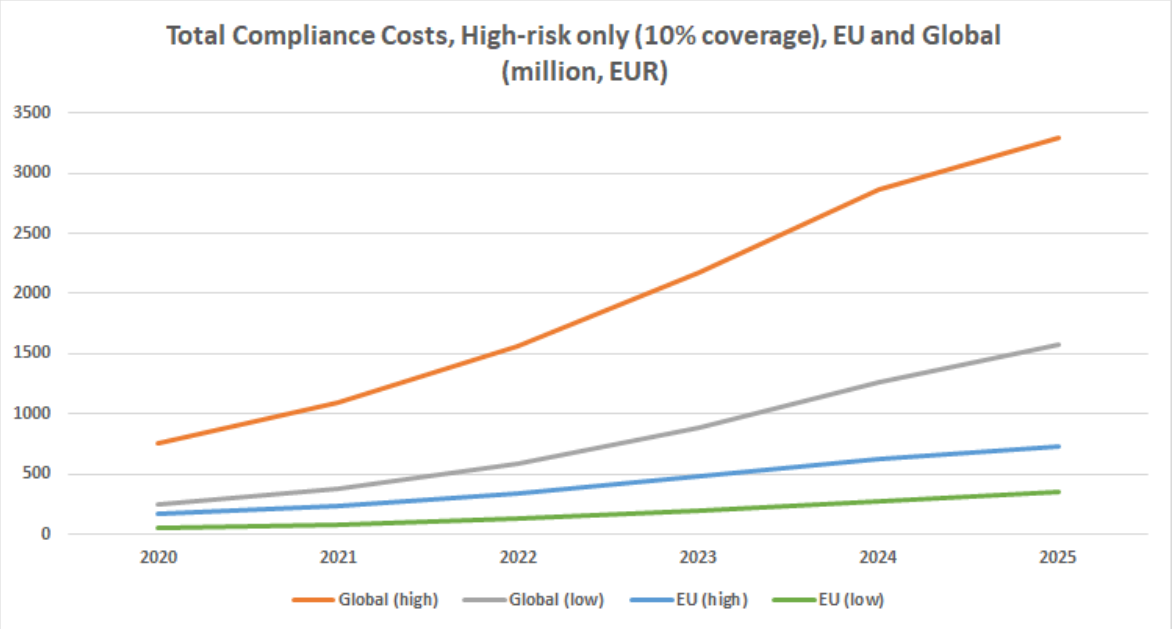
economy is expected to spend EUR 131 million - EUR 345 million on compliance, while the cost to the global economy is expected to range from EUR 593 million to EUR 1.569 billion. In 2025, compliance may cost the EU economy EUR 347 million – EUR 726 million, and the global economy EUR 1.578 billion – EUR 3.3 billion. While these estimates assume that high-risk AI systems only count for 10% of total AI investments, the actual proportion is unknown and will depend on the definition of high-risk AI systems. The private sector will also respond to the new regulation and thus the equilibrium high-risk AI investment will be determined endogenously.

Table 31 - Compliance cost of regulation on only 10% of AI systems (EUR million)

10% coverage	2020	2021	2022	2023	2024	2025
Global (high)	761.09	1,095.15	1,568.86	2,176.22	2,863.33	3,300.32
Global (low)	246.21	383.07	593.36	889.98	1,266.17	1,578.04
EU (high)	167.44	240.93	345.15	478.77	629.93	726.07
EU (low)	54.17	84.27	130.54	195.80	278.56	347.17

Source: Authors' computation

Figure 21 - Total compliance cost of high-risk only regulation



Source: Authors' computation

5. Cost estimation of conformity assessment procedure

Compliance with the requirements is only the first part of the process. The next step is to apply for certification before the AI system can enter the EU market. The process of certification usually requires an independent third-party - a notified body - to verify that the AI system is compliant with the requirements. This process is usually called conformity assessment.

Similar to the certification scheme under the Medical Device Regulation (MDR), the AI regulation and the associated certification process follow two different procedures, with details usually listed in annexes to the regulation. This cost estimation exercise attempts to mirror Annex IX (Quality Management System) and Annex X, together with Annex XI(A) (Type Examination and Internal Production Control).

The two procedures are substantially different in their cost implications for the applicant. Annex IX of the MDR requires the set-up of a Quality Management System (QMS) within the firm, which incurs a large one-off cost but is then easily subject to regular updates of standards. The notified body usually performs an on-site audit of the QMS and will review the technical documentation for each product. The notified body will also review the QMS once a year, as part of continuous monitoring.

Annex X and Annex XI(A) of the MDR define a conformity assessment for a single product. Under Annex X, the firm prepares the technical documentation to prove that the product complies with the requirements, which are then reviewed by a notified body. The notified body will test the product and verify the information given in the technical file. Annex XI(A) requires an audit of the production QMS.

This cost estimation exercise consists of two sections. First, an estimate of the costs of a conformity assessment of a single product under the procedure of EU-type examination. Again, the AI product is assumed to be a unit of AI (a development cost of EUR 170,000). This report applies both a bottom-up approach and benchmarking in reaching the cost estimates. Second, an estimate of the costs of a conformity assessment of a single product (a unit of AI) under a QMS. Under this procedure, the cost also includes the one-off expense of setting up a QMS (including benchmarking and validation from different experts).

In practice, the costs could vary substantially based on a range of unknown factors. For instance, a piece of hardware may be connected to more than one AI system. This report assumes that the reference product is embedded with one AI unit that contains one AI system, or the product itself is the AI system that costs EUR 170,000 to develop.

a. EU-type examination

Many types of products have been regulated by directives and regulations that require certification, such as toys, medical devices, machinery, etc. Other products are subject to a General Product Safety Directive that requires no conformity assessment by an independent body. In other words, some products have already been subject to existing conformity assessment while others interacting with an AI system will have to undertake a new conformity assessment procedure. Two scenarios are considered: 1) an AI assessment in addition to an existing conformity assessment; and 2) a new AI conformity assessment.

The report considers the possibility that a final product manufacturer/applicant embeds an AI component that is developed by an upstream producer. It is possible that the applicant has not provided additional training data in the production process and thus does not own the training data or records and may face difficulties in providing such information. The expert advice provided through interviews as part of this study suggested, however, that an advanced AI product likely involves additional data for tailoring the AI system to fit the intended use of the product. In this case, the product producer would need to collaborate with the upstream producer to prepare the technical documentation. Any certification of the upstream AI product may help to improve the credibility of the documentation, but the material benefit is believed to be minimal. The report thus omits this particular possibility and focuses instead on cases where the applicant is required to provide information on compliance with all requirements of the proposed regulation.

Bottom-up approach

Case 1: New conformity assessment

For conformity assessment applicants (‘applicants’), the preparation of technical documentation is the main in-house cost. That preparation involves both compliance costs (see earlier sections for estimates) and actual information documentation costs. Technical

documentation entails the translation or summary of existing internal information into documents for certification.

The reference cost level per hour for an assessment is taken from various fee quotations offered by notified bodies specialising in conducting conformity assessment under the MDR. Reviewing documentation can cost approximately EUR 400 per hour, while an on-site audit may cost EUR 300 per hour (e.g. Tuvsud, n.d.).

The applicant prepares the necessary technical documentation. For consistency, the study assumes that the hourly wage per labour in-house is EUR 32. The time needed to prepare all documentation varies across products, but the time taken to prepare the documentation for a laptop is 20 days, or 150 hours (European Commission, 2014a, p.49). Taking this number of hours as a reference, the total in-house cost for technical documentation for applicants amounts to EUR 4,800 (see Table 34). An expert in the field of medical device manufacturing commented in an interview that preparing the technical documentation of a product produced by a SME may amount to anything between EUR 10,000 to EUR 30,000, including internal testing costs (taken into account in the calculation of compliance costs). The estimate used here – EUR 4,800 – relates solely to paperwork and is believed to be within a reasonable range.

Mirroring the MDR, a notified body would then review the technical documentation and perform audit and testing. Concerning training data and record-keeping, the notified body is supposed to conduct a sufficiently broad audit to verify if the information provided in the documentation is true and fair. On outcome-related requirements, namely, human oversight and robustness and accuracy, the notified body may conduct additional testing to evaluate the AI system. An interview with an experienced manager of regulations and standards revealed that it may take two days to review the documentation and another five days of on-site testing for an average product, which amounts to more than 50 hours of work. The time needed to assess a more complex system 'could easily double or triple the time'. As the use of 'AI unit as the unit of analysis accounts for the complexity of the AI system, 20 hours is taken for review and 33 hours of audit and testing. In total, the cost to the notified body is EUR 18,200, which it is expected to pass to the applicant. As a result, the applicant pays EUR 23,000 for certification of one AI system, 13.5% of the compliance costs. An integrated AI product may consist of many AI systems. For ease of understanding, the time needed to complete a task is expressed in hours.

Table 32 - Reference cost per hour

Task	Cost per hour (EUR)
Review	400
Testing	400
Audit	300
In-house labour, per hour	32

Table 33 - Conformity cost estimates of Case 1

	Developing technical file (hour)	In-house cost (EUR)	Review of technical documenta-tion (hour)	Testing (hour)	Audit (hour)	Total (hour)	Total cost to notified body (EUR)	Total cost (EUR)
Training data	30	960	4		10	14	4,000	
Record-keeping	30	960	4		4	8	2,800	

Information provision	30	960	2			2	400	
Human oversight	30	960	2	4		6	1,600	
Robustness and accuracy	30	960	4	15		19	5,600	
Total cost (EUR)		**4,800**					**18,200**	**23,000**

Case 2: New conformity assessment

In this case, the AI-embedded product must pass an existing conformity assessment by a notified body, and the AI system is tested within the same procedure. The cost of the technical documentation preparation is shared with the existing conformity assessment. For instance, when the applicant documents the information on the product's functionality, the AI system would be described together with other details on the hardware. Given the complexity of an average AI system, the study team reduced the cost of technical documentation by half. Even if the cost of documentation preparation is reduced, however, the AI system must be sufficiently described and explained, requiring the same amount of information to be provided to the notified body. The notified body will therefore likely need the same amount of time to review the documentation as in Case 1 above. The same logic was applied to the audit of data and records. However, testing of the product, hardware and software, could be conducted at the same time, saving an estimated half of the time and cost. The in-house cost is now EUR 2,400 and the fee paid to the notified body is EUR 14,400, costing the applicant a total of EUR 16,800. This is only the cost of the AI conformity assessment, however - the total cost of the whole conformity assessment is the sum of the AI conformity assessment and the existing conformity assessment of the product.

Table 34 - Conformity cost estimates of Case 2

	Developing technical file (hour)	In-house cost (EUR)	Review of technical documentation (hour)	Testing (hour)	Audit (hour)	Total minutes (hour)	Total cost to notified body (EUR)	
Training data	15	480	4		10	14	4,000	
Record-keeping	15	480	4		4	8	2,800	
Information provision	15	480	2			2	400	
Human oversight	15	480	2	4		6	1,600	
Robustness and accuracy	15	480	4	15		19	5,600	
Total cost (EUR)		**2,400**					**14,400**	**16,800**

Benchmark

A benchmarking estimate allowed these study estimates of the cost of an EU-type examination to be compared with the conformity assessment costs of other legislation, thereby verifying that they are reasonable.

The **first benchmarking estimation** focuses on national IT security certification schemes, the costs of which are presented in the European Commission's Impact Assessment

accompanying the proposal for the 'Cybersecurity Act' (European Commission, 2017). The costs of assessing such systems for conformity with regulatory requirements could well resemble those of assessing high-risk AI, as both system types require high technical skills and consider similar risks, e.g. in terms of safety.

In France, the *Certification Sécuritaire de Premier Niveau* (CSPN) is a quick and cheap IT security certification scheme compared to the common criteria approach. The cost of each CSPN certification is about EUR 25,000-EUR 35,000 and takes around three months. Another example comes from the Netherlands, where the Baseline Security Product Assessment (BSPA) has been created to assess the suitability of IT security products for use in the 'sensitive but unclassified' domain. This certification costs on average EUR 40,000 and takes up to two months.

There may be more dimensions to the conformity assessment for a product embedding an AI system or for a stand-alone AI system. However, these overall estimates are considered reasonable as an average value for the entire market.

The **second benchmarking estimation** is derived from the case studies in the Evaluation of Internal Market Legislation for Industrial Products (European Commission, 2014b). This analysis is conducted by the Centre for Strategy and Evaluation Services (CSES) for the European Commission and features costs for implementing the applicable internal market legislation for 10 different products: electric motors, laptops, domestic refrigerators and freezers, lifts, gardening equipment, fuel dispensers (measuring instruments), air conditioners, integrated circuits, snow-ski footwear and bicycles. To compare the costs of conformity assessment procedures between these products and those embedding an AI component, the analysis focused on four products: laptops, gardening equipment, lifts, and air conditioners. These were chosen because the data available were the most comprehensive in quality and quantity. These products also represent a relatively high level of innovation and technological change, and the market actors include both large dominant multinationals and SMEs.

Each case study consists of interviews with companies and industrial associations, in which the study team asked them about the costs they faced for different actions deriving from the new legislation. The conformity assessment formed the bulk of the interview, which included: relevant testing and development of the technical file, use of notified body if/when required, preparation of conformity and CE marking.

Table 35 - Summary of case studies (cost in EUR)

Case study	Technical file preparation	Review by notified body	Testing	Conformity to type	Total Cost (EUR)
Laptops	4,800	15,000	5,000	Negligible	24,800
Gardening equipment	2,100	4,000	100 – 1,000,000	130	6,230+
Lifts	BAU	25,000 – 30,000	200 – 1,000	Negligible	25,200 – 31,000
Air conditioners (per year, multiple products)	106,169	74,880	53,653	Included in technical file cost	234,702

Source: Evaluation of Internal Market Legislation for Industrial Products (European Commission, 2014b)

The costs in the table represent averages of the information shared by the different entities interviewed. Information was not available on the cost of conformity assessment for air

conditioners, with interviewees mentioning that each firm could produce up to five different types of air conditioners.

The total costs for each product ranged **from EUR 6,230 to around EUR 40,000** and are therefore consistent with the study's estimates and with the first part of the benchmark. There are still minor differences between each product for the allocation of costs among the different phases: for lifts, for instance, most of the costs come from the review by the notified body itself, while the documentation and testing costs are negligible. In turn, this cost does not apply to air conditioners, where the preparation of the technical file represents almost half of the total costs.

The **third benchmarking estimation** includes a structured collection of sources from notified bodies and other impact assessment studies for EU regulations. Notified bodies' fees were used for the hourly cost of labour, while the number of hours needed for each phase was benchmarked from diverse legislation. These time estimates will likely differ for AI products but are nevertheless useful to compute an average benchmark estimate of conformity assessment costs.

Table 36 - Cost estimates using benchmark averages (cost in EUR)

	Technical file preparation	Review by notified body	Testing	Conformity to type	External audit	Total cost	Total with external audit
Time (hour)	97,5	15		15	21		
Hourly rate (EUR)	32	400		32	300		
Cost (EUR)	3,120	6,000	10,000	480	6300	**19,600**	**25,900**

The hourly rate for internal labour is again set at EUR 32 (Eurostat NACE Rev2 sector J (information and communication)). The study team conducted a high-level workshop on 7 October to discuss these estimates with field experts, with the majority noting that this labour cost was too low and not representative of the salaries of AI experts. The other labour costs for review and audit are averages of multiple sources (quotation fees from TÜV, BSI, UKAS, etc.).

For testing, the benchmark estimate is an overall number that is not derived from time and labour costs, as its cost depends on the infrastructure required and available. From the benchmarking, the cost of testing could go from EUR 100 to EUR 1,000,000. The study team assumed an average company that had already carried out testing as BAU and owns the testing infrastructure and thus only needs to conduct further testing to prove full conformity during the notified body review – the average cost of testing was thus set at EUR 10,000. However, the experts did not validate this assumption at the workshop, instead suggesting that testing costs might be higher, as no common procedures are in place for testing AI systems and products. This might entail high additional costs for companies, at least in the initial years after the regulation comes into force.

In conclusion, the benchmark estimations facilitated a comparison of the costs of other conformity assessments with the study estimates and gathered feedback from stakeholders (through a workshop and interviews) on the reasonableness of those estimates.

b. Full quality assurance

The procedure of full quality assurance (QA) refers to setting up an internal production QMS, audited by a notified body, together with a review of technical documentation of each product. In theory, applicants are not required to attain international standards (e.g. ISO and IEC) to fulfil the requirements of the regulation. In practice, however, companies often acquire international standards as additional supports for their product conformity assessments by notified bodies. Each international standard has different requirements that may incur substantial – and primarily one-off – costs. Money and time are spent on preparing administrative documents and adjusting internal production processes, but once the system is certified, the additional cost of certifying another product is relatively low.

Expert advice in the medical device industry indicates that most of the established companies are equipped with a QMS and thus prefer a conformity assessment procedure based on that system. In some sectors, internal quality systems are de facto required to get market access. Under the MDR, because the vast majority of software is unclassified under the Regulation, a lot of software-only companies now face the costs of certifying their QMS with ISO 13485 for the first time, and having the conformity assessment of their technical file done by a notified body. The study thus assumes that even for companies with a QMS already in place, additional costs will arise from upgrading that QMS to meet new regulatory requirements. In addition, companies that already have an up-to-date QMS will still have to face the costs of regular audit and preparation of technical documentation. The expert stakeholders explained that in the medical device sector, notified bodies do not offer software type examination (contrary to hardware type testing) following Annex X of EU MDR, generally because of time, cost and lack of specific expertise. Expert industry stakeholders explained that software type-testing was not seen as a viable option.

Short surveys were used as part of the study's high-level workshop to gather structured and written feedback from all participants. On the costs of conformity assessment procedures, participants were asked which of the procedures companies would prefer if they had the choice between type examination and full QA. Of the 11 respondents, five stated that full QA would be preferred, three answered 'others' (e.g. self-certification), and three did not answer. The majority of respondents (five) felt that type examination would be the most costly procedure, with some noting that it is too bureaucratic and would therefore be more costly. By contrast, the full QA procedure would be more manageable and flexible, as well as sufficiently robust in case of product updates, i.e more realistic due to frequent technology enhancements.

The upfront cost of this procedure may be too high for SMEs, however, and could effectively impose an entry barrier to the market. Some participants at the workshop suggested sharing platforms that would allow SMEs to share the costs of QMS in order to stay competitive. Others referred to the practice of subsidising testing for SMEs, as done in Singapore.

To estimate more precisely the cost of this conformity assessment procedure, two in-depth interviews were carried out with Koen Cobbaert, a distinguished expert specialising in health software at Philips. Mr Cobbaert shared his extensive knowledge about the cost of conformity assessments under the MDR, based on his experience. This information proved very useful in assessing the costs of the same procedure for AI products, as the medical devices sector already features some AI software and requires a similar level of high technological expertise. The risks in case of error or malfunctioning can be expected to have the same magnitude concerning safety and fundamental rights.

Mr Cobbaert outlined a case where one SME with 100 employees launches one medical device on the EU market. His estimates were then used to gather structured feedback from the participants at the high-level workshop. Participants generally agreed with them: of 11 respondents, six believed these to be the most realistic estimates. Several noted that these estimates include the costs of setting up new systems, and that the costs would be lower

once the systems were set up. The section below distinguishes between two cases: 1) where the SME already used a QMS prior to the AI legislation, and 2) where the SME does not have a QMS in place.

One-off costs of QMS

The first phase of this procedure is setting up a QMS that is compliant with the regulation. For now, it can be supposed that the QMS will have to be compliant with ISO9001 (the international standard that specifies requirements for a QMS) and with the overall AI regulation. It is likely that a standard specific to AI quality management system will be published, like that for medical devices (IEC13485). One standard on AI was already published in early 2020 by the Joint Committee ISO/IEC JTC1/SC 42, working on standardisation in the area of AI, on topics related to the trustworthiness of AI (ISO, n.d.). Experts in the area highlighted that every standardisation organisation has begun to tackle AI, but that the overall process will take several years to be complete. **The cost of setting up a QMS is estimated to range between EUR 80,000 and EUR 160,000**, including the human resources to set up the processes. The variance in the cost is related to the complexity of the organisation and the need to hire external consultants. Around EUR 100 can be added for the purchase of a standard.

This cost will not apply to a company with some kind of QMS in place. However, there might still be additional costs incurred from analysing and interpreting the regulation, conducting a literature study to look for state-of-the-art practices and existing standards, and updating and upgrading the QMS accordingly. Based on the EUR 32 hourly rate, **the cost for an SME will amount to EUR 5,280.**

Once the QMS is set-up or upgraded, employees will then have to be trained. For a new QMS, 30 minutes of training per employee is assumed, with 20 minutes for an upgraded system. With 100 employees and a EUR 32 hourly rate, **staff costs amount to EUR 1,600 for a new QMS and EUR1,070 for updating an existing QMS.**

The company will also have to draw up documentation on the QMS in order to allow for consistent interpretation by the third party. Such documentation is assumed to be drawn up during the set-up of the system itself, thus its cost is reflected in the given range. There will be an additional 100 hours of FTE needed to compile evidence, make all documents consistent and coherent, ensure that they exhaustively cover compliance, and write the narrative to be understandable by a third party. Again at the EUR 32 hourly rate, **documentation costs amount to EUR 3,200.** This cost could substantially increase if the company uses external counsel.

Whether new or upgraded, the QMS has to be audited by the notified body and proven compliant with the standards and the regulation. This audit costs EUR 1,550 per day and the number of days will depend on the specific structure and complexity of the company. **The overall cost of the audit could reach EUR 65,100.**

Setting up an information security management system (ISMS) is likely to become the best practice for complying with the AI regulation, as it is in the medical devices sector. The need for ethical technology assessment is evident, especially for high-risk AI applications. The cost for setting up such a system amounts to EUR 30,000, including labour costs, and will likely be less if no personal data are processed by the device.

The ISMS will also need to be audited by a third party to prove compliance with regulation and standards. This is expected to cost EUR 32,550, but again could be less in the absence of personal data involved.

One-off cost for individual products

For each product, the company will have to compile documentation to prove that the product complies with the AI regulation. This includes the time for developers and others to write the documentation and the substantiating pieces of evidence, as well as the time to make it compliant and readable from a regulatory perspective. This amounts to between EUR 10,000 and EUR 30,000, depending on the complexity of the product and of the organisation.

The notified body will then have to review this documentation to ensure that it is compliant with the requirements. At a EUR 400 hourly rate and with the review expected to take between one and two-and-a-half days, the cost will **range from EUR 3,000-EUR 7,500 for the notified body to monitor compliance with the documentation requirements.**

Ongoing costs

Remaining compliant with the regulation and standards will require the company to undertake yearly audits of its QMS. With an hourly rate of external audit of €300, **the costs for yearly audits by a notified body could go up to roughly €9,000 per year** (involving two people coming for two days)**.**

To stay compliant over the lifecycle of the products - and thus prepare for the annual audits - the QMS should be continually monitored and improved. Standards are updated over time and the company needs to make sure that its processes are continually updated to reflect those improvements in standards. This is likely to be especially true for AI, which is changing substantially and rapidly. Overall, this requires oversight and maintenance of the system, as well as regular updates of the technical documentation. For companies with a QMS in place prior to AI regulation, this cost is indistinguishable from BAU and best practice. Companies building a QMS for the first time, however, will have to hire someone to take charge of this (mirroring the person responsible for regulatory compliance in the MDR) or outsource it. This cost is estimated by considering **one FTE over the year, which amounts to EUR 62,400** (EUR 32 hourly rate, 7.5 hours per day, 260 days per year).

Table 37 - Cost of conformity assessment using full QA procedure

	Company using QMS	Company not using QMS
One-off costs for QMS (EUR)		
1- Set-up/upgrade QMS	80,100-160,100	5,380
2- Training	1,600	1,070
3- Audit of QMS	32,550-65,100	32,550-65,100
4- Set-up ISMS	30,000	30,000
5- Audit of ISMS	32,550	32,550
6- Draw up documentation	3,200	3,200

One-off costs for products (EUR)		
7- Draw up documentation	10,000-30,000	10,000-30,000
8- Review by notified body	3,000-7,500	3,000-7,500
Total one-off costs	**193,000-330,050**	**117,750-174,800**
Ongoing costs (EUR)		
7- Annual audit	9,000	9,000
8- Oversight	62,400	62,400
Total ongoing costs per year	**71,400**	**71,400**

The set-up of a QMS and the conformity assessment process for one AI product is estimated to cost between EUR 193,000 and EUR 330,050. An estimated additional yearly cost of EUR 71,400 will also be borne by the company to maintain compliance over time.

For a company that has to upgrade its QMS to comply with the additional requirements of the AI legislation, the overall cost is estimated to be between EUR 117,750 and EUR 174,800. An estimated additional yearly cost of EUR 71,400 will also be borne by the company to maintain compliance over time.

Some limitations to these cost estimates remain, as some specificities of AI and particular costs could not be captured. Firstly, the cost of building and auditing the QMS and ISMS could increase substantially if the development process of the products is fractured between different sites. Additional costs may also derive from the staff time invested in responding to criticism from notified bodies and correcting the issues highlighted. This cost relies on the complexity of the products and also on each company: it assumes that having a certified full QA procedure would reduce the risks for companies of having negative feedback from the notified body. Finally, the study does not strictly monetise a major opportunity cost – that of delayed market entry. It must be acknowledged that companies will have to wait for certification before launching their product on the EU market and that this comes at a cost.

c. Cost estimates for smaller enterprises

For a perspective on smaller enterprises, the numbers provided by Mr Cobbaert are used, with fixed costs distinguished from variable costs, and estimates dependent on number of staff. While training and documentation costs are roughly proportionate to the size of the enterprise or the complexity of the AI system, the set-up and audit costs are mixed. 50% of those set-up and audit costs are assumed to be fixed, i.e. they are unavoidable once the enterprise decides to enter the regulated market. The other half of the costs is proportionate to the number of staff, which is a simple proxy for the complexity of the AI system. Other than this classification of costs, Mr Cobbaert's numbers for an SME employing 100 employees are maintained.

According to Mr Cobbaert, for an enterprise employing 100 employees and having no existing QMS, setting up and maintaining a QMS, plus examining the product, ranges from EUR 264,400 to EUR 401,450. The cost is calculated at each size (0-100 employees) of an enterprise as the following:

$$TotalCost = FixedCost + VariableCost \times \frac{Size}{100}$$

Therefore, a hypothetical enterprise that employs no staff would pay only the fixed cost and an enterprise having 100 employees would pay the total amount suggested by Mr Cobbaert.

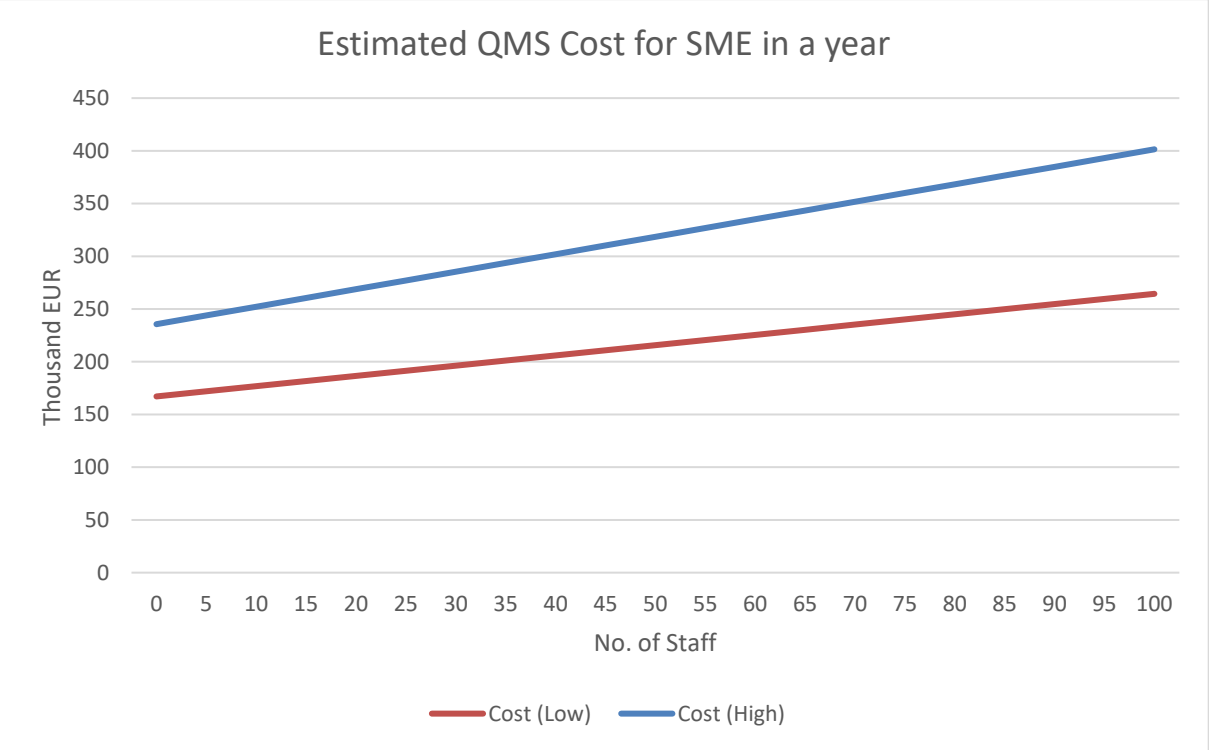
Set-up and audit costs are mixed because while they may highly correlate with the complexity of the production, some fixed overhead costs are unavoidable. However, the annual audit may involve some standard procedures that are identical, irrespective of the complexity of the AI product. The audit cost is thus set at 100% fixed.

Table 39 shows the classification of costs into either fixed or variable costs, or both. Mixed costs are split 50-50. Figure 23 plots the estimated total cost against the number of employees**. For an enterprise of 50 employees, the total cost ranges from EUR 216,000 to EUR 319,000 for one product in year one.**

Table 38 - Fixed and variable costs for enterprises without a QMS in place

	Enterprises without a QMS	
	Fixed costs (EUR)	**Variable costs (EUR)**
One-off costs for QMS (EUR)		
1- Set up QMS	40,050-80,050	40,050-80,050
2- Training		1600
3- Audit of QMS	16,275-32,550	16,275-32,550
4- Set up SMS	15,000	15,000
5- Audit of SMS	16,275	16,275
6- Draw up documentation	1,600	1,600
One-off costs for products (EUR)		
7- Draw up documentation	5,000-15,000	5,000-15,000
8- Review by notified body	1,500-3,750	1,500-3,750
Ongoing costs (EUR)		
7- Annual audit	9,000	
8- Oversight	62,400	

Figure 22 - Estimated Costs of Entering the Market by Setting up a QMS



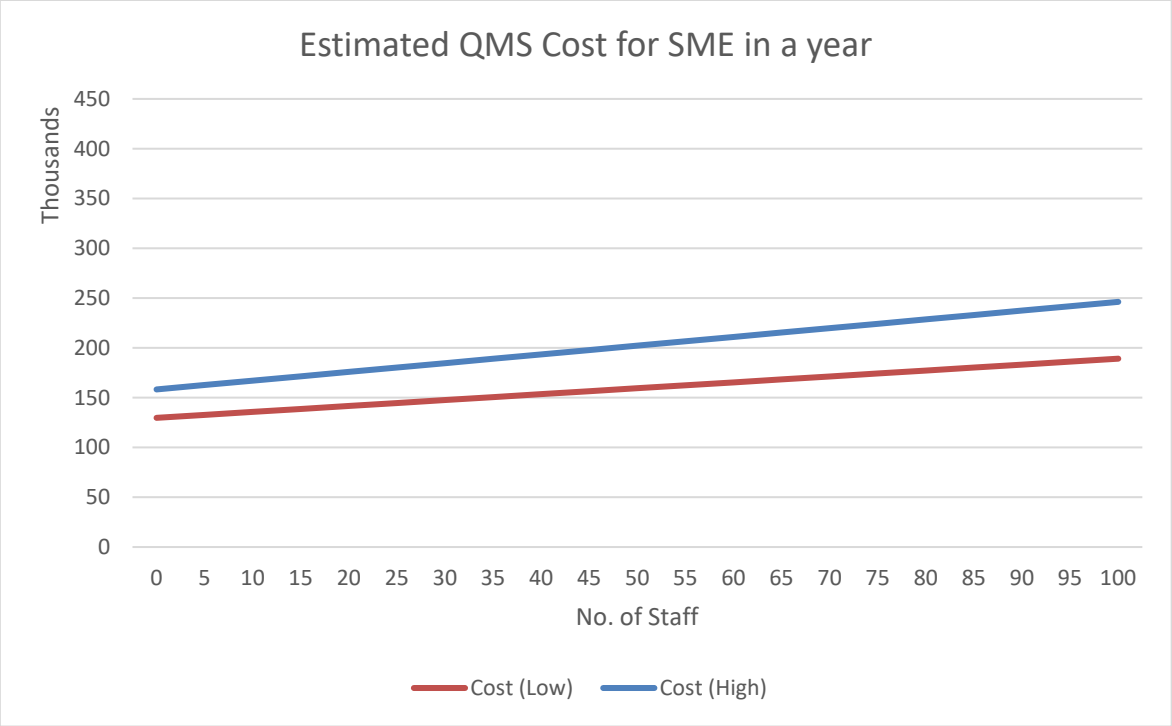
The case is similar for enterprises with a QMS in place. Assessing an AI system, even when provided with clear requirements, will be unfamiliar to both AI producers and notified bodies, auditing the QMS may incur roughly the same cost as if the enterprise had no QMS in place. The range of audit costs of QMS above already covered the possibility that auditing an upgraded system is less costly. **An enterprise employing 50 persons would pay roughly EUR 159,000-EUR 202,000 for upgrading and maintaining the QMS, and bringing one AI product to market.**

Table 39 - Fixed and variable costs for enterprises with a QMS in place

	Enterprise with QMS	
	Fixed costs (EUR)	**Variable costs (EUR)**
One-off costs for QMS (EUR)		
1- Upgrade QMS	2,690	2,690
2- Training		1070
3- Audit of QMS	16,275-32,550	16,275-32,550
4- Set up SMS	15,000	15,000
5- Audit of SMS	16,275	16,275
6- Draw up documentation	1,600	1,600

One-off costs for products (EUR)		
7- Draw up documentation	5,000-15,000	5,000-15,000
8- Review by notified body	1,500-3,750	1,500-3,750
Ongoing costs (EUR)		
7- Yearly audit	9,000	
8- Oversight	62,400	

Figure 23 - Estimated costs of entering the market by upgrading an existing QMS



The propensity of SMEs to invest in high-risk - and thus likely regulated - AI systems should also be taken into account. Some stakeholders interviewed stated that they might refrain from producing any regulated AI systems to avoid additional costs. **A market force that would shift investment away from the regulated market implies a lower global compliance cost.** SMEs may lack significant funds and thus choose to stay away from the regulated market.

6. Adding compliance costs and conformity assessment costs

System providers could choose to apply certification through two procedures. Without a proper estimate of the use of the two procedures, it is impossible to scale the cost of conformity assessment to the population. The approximation here makes several assumptions:

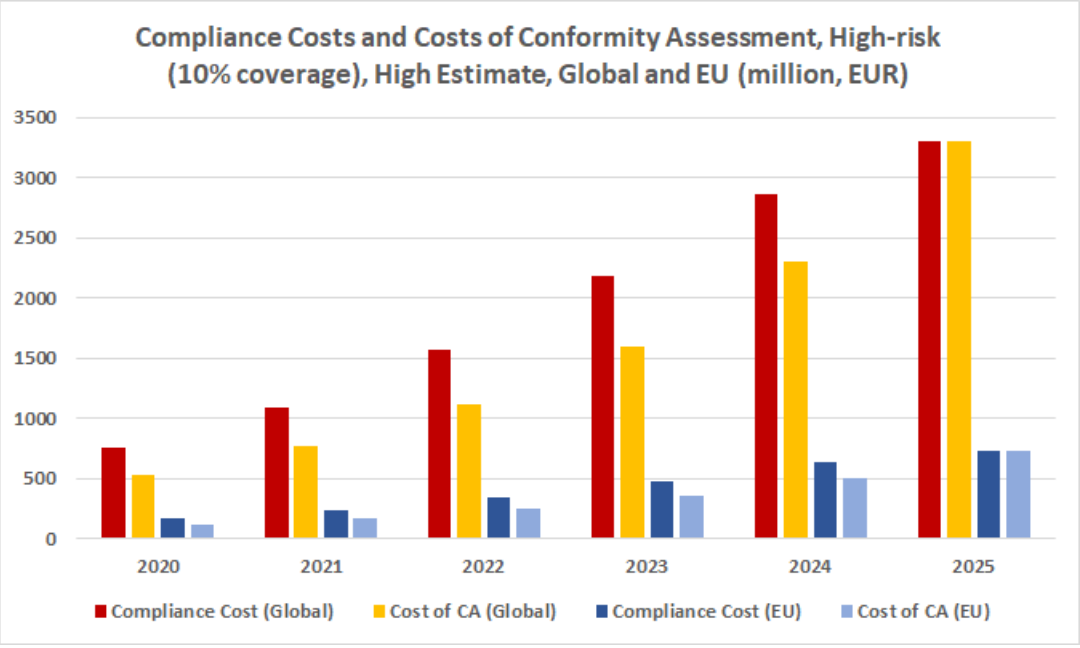
- All AI units will pass through the EU-type examination;

- 70% of AI products will fall under existing legislation that requires a conformity assessment (the large majority of manufacturing goods, whether produced in or imported to the EU, are subject to some sorts of conformity assessment);
- The regulation will concern only high-risk AI systems;
- AI investment follows the upper-bound estimates.
- In 2022, with the BAU factor taken into account, the sum of costs of compliance and conformity assessment to the EU economy is expected to reach EUR 589 million and EUR 2.7 billion to the global economy. In 2025, the cost to the EU and the global economy would rise to EUR 1.5 billion and EUR 6.6 billion, respectively.

Table 40 - Costs of compliance and conformity assessment, BAU considered (EUR million)

	2020	2021	2022	2023	2024	2025
Europe						
Compliance	167	241	345	479	630	726
Conformity assessment	118	170	244	351	505	727
Total costs	**285**	**411**	**589**	**830**	**1,135**	**1,453**
World						
Compliance	761	1,095	1,569	2,176	2,863	3,300
Conformity assessment	536	771	1,109	1,596	2,297	3,306
Total costs	**1,297**	**1,866**	**2,678**	**3,773**	**5,161**	**6,606**

Figure 24 - Costs of compliance and conformity assessment to the EU and global economies



7. Other costs

a. Registration cost

Registration cost is believed to be trivial compared to compliance and conformity costs. Taking the MDR as an example, the applicant (manufacturer/authorised representative/importer) is required to apply to a national authority. A Single Registration Number (SRN) will be issued upon validation and any related information and data can be submitted to the EUDAMED database, which is a multipurpose open platform for registration, notification and dissemination. Table 42 lists several examples of registration fees. While AI products may involve some additional expertise on the part of the national authorities, the registration fee is still low. Based on the benchmarking, this report estimates a registration fee of EUR 200 per AI product.

Table 41 - Examples of registration costs

Regulation	Country	Registration Fee	Remarks
MDR	UK	GBP 100 per submission[79]	One submission could include multiple products of the same code
	Ireland	EUR 140 per registration[80]	
	Switzerland	CHF 200 per hour of work[81]	
	Austria	Free of charge	
	Denmark	DKK 1,159	Plus annual fee depending on product type/company size
Energy Labelling Regulation		Free of charge[82]	
Fertilising Product Regulation	Finland	EUR 85[83]	

b. Other costs: AI Board

Among the costs generated by the prospective regulation on AI is that of setting up an AI Board as part of the EU institutions.

[79] UK government guidance available at: https://www.gov.uk/guidance/register-as-a-manufacturer-to-sell-medical-devices

[80] http://www.hpra.ie/homepage/medical-devices/registration

[81] Swiss advice available at: https://www.swissmedic.ch/dam/swissmedic/en/dokumente/medizinprodukte/mepv/bw630_10_002_d_mb_srn-faq.pdf.download.pdf/BW630_10_002e_MB_SRN_FAQ.pdf

[82] European Commission product database available at: https://ec.europa.eu/info/energy-climate-change-environment/standards-tools-and-labels/products-labelling-rules-and-requirements/energy-label-and-ecodesign/product-database_en

[83] Finnish advice available at: https://www.ruokavirasto.fi/en/companies/feed-and-fertiliser-sectors/fertilizer-sector/

European Data Protection Board

A useful benchmark in this respect is the European Data Protection Board (EDPB), an EU body in charge of the application of the GDPR. The EDPB is composed of the head of each national data protection authority and the EDPS (or their representatives). The EDPB helps to 'ensure that the data protection law is applied consistently across the EU' and works to ensure effective cooperation among data protection authorities. The Board will 'issues guidelines on the interpretation of core concepts of the GDPR' and is also called on to issue binding decisions on disputes regarding cross-border processing, thereby ensuring uniform application of EU rules.[84] The EDPB:

- Provides general guidance (including guidelines, recommendations and best practice) to clarify the GDPR;
- Adopts consistent findings, designed to make sure that the GDPR is interpreted consistently by all national regulatory bodies, for example in cases relating to two2 or more countries;
- Advises the European Commission on data protection issues and any proposed EU legislation of particular importance for the protection of personal data;
- Encourages national data protection authorities to work together and share information and best practices.[85]

In 2018, the EDPS was allocated a budget of EUR 14,449,068. Title I of the EDPS budget comprises five articles and is designed to cover expenditure relating directly to the members and staff of the institution. The amounts entered in Title I of the budget for staff came to a total of EUR 7,223,575. The utilisation rate for the appropriations entered in Title I was 96.11% of the committed amount, totalling EUR 6,942,838. The EDPB currently has 21 employees.

European Medicines Agency

Another possible benchmark is the European Medicines Agency (EMA), which counts 869 employees and had an annual budget of EUR 358.1 million in 2020. The EMA was set up in 1995, with funding from the EU and the pharmaceutical industry, as well as indirect subsidies from Member States. Its stated intention is to harmonise the existing work of national medicine regulatory bodies. It underlines four main objectives:

- Facilitate development and access to medicines;
- Evaluate applications for marketing authorization;
- Monitor the safety of medicines across their lifecycle;
- Provide information to healthcare professionals and patients.[86]

Around 86% of the EMA's budget derives from fees and charges, 14% from the EU contribution for public health issues, and less than 1% from other sources.

[84] See https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/enforcement-and-sanctions/enforcement/what-european-data-protection-board-edpb_en

[85] See https://edpb.europa.eu/our-work-tools/general-guidance/gdpr-guidelines-recommendations-best-practices_en

[86] See https://www.ema.europa.eu/en/about-us/what-we-do

Of the total budget in 2021:

- Approximately EUR 330.4 million will come from fees and charges levied for regulatory services;
- Approximately EUR 55.4 million is expected in income from the EU, mainly to support the policies for orphan and paediatric medicines, advanced therapies, micro-enterprises and SMEs.

The EMA charges a fee for processing applications from companies that want to bring a medicine to the market. It also charges fees for services related to the marketing of medicines in the EU in areas such as scientific advice, inspection and the establishment of maximum residue limits (EMA, n.d.).

The EMA coordinates the scientific evaluation of applications and related work with the national medicines regulatory authorities in the Member States. It compensates the national authorities for this work and the involvement of their staff members in its scientific committees, working groups and other activities.

In 2021, it is estimated that EUR 134 million will be paid to the national medicines regulatory agencies from the budget [87]

European Chemicals Agency

The European Chemicals Agency (ECHA) is the EU agency that manages the technical and administrative aspects of the implementation of the EU Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) Regulation. The ECHA collected a total of EUR 31,273,450 (in 2018: EUR 78,208,000) in registration fees, EUR 5,100,000 in authorisations, EUR 21,525 from appeals, EUR 168,000 from classification, labelling and packaging, and EUR 1,000,000 in other fees and charges. The ECHA's annual revenue was EUR 112,283,950, and annual expenditure was EUR 113,448,180 (ECHA, 2019; ECHA, 2020).

European Union Agency for Cybersecurity

A supporting agency for EU policy on network and information security is the European Union Agency for Cybersecurity (ENISA). ENISA conducts candidate certification schemes within an EU cybersecurity certification framework. It engages with public services, as well as with industry and standardisation organisations for the certification of ICT products and services, within the meaning of the Cybersecurity Act. ENISA's annual revenue in 2018 was EUR 11,425,705.72 (ENISA, 2019). In 2019, ENISA's annual voted budget was EUR 16,932,952.05.

8. Cost impact on SMEs

During the public consultation, some stakeholders warned of the costs that would be incurred by a mandatory compliance and conformity assessment for SMEs. Following a bottom-up approach and including benchmark values, the study attempts to assess the cost incurred for an average firm developing an average AI product. It is acknowledged that these costs will vary according to the size or age of the firm and complexity of the product,

[87] See https://www.ema.europa.eu/en/about-us/how-we-work/governance-documents/funding

with the benchmark comparison and in-depth expert interviews used to gain insights into the estimated cost impact for SMEs.

SMEs are different from large corporations in that they are often less prepared for new regulations. While large corporations have been following the development of other regulations with dedicated human resources, SMEs may have to invest additional resources in their operation and management in-house to ensure compliance with a new regulation. Even if in-house investments could be avoided, they may have to pay for services from consultancy firms or notified bodies. In addition, overhead costs (including legal fees) are proportionally more costly for SMEs, which do not enjoy sufficient economies of scale. As a result, the new regulation may impact heavily on SMEs' profit margins. It is also true, however, that their products might be less sophisticated and thus require fewer resources to ensure compliance compared to large corporations. Given the differences foreseen, the cost impacts on SMEs are analysed and compared as outlined below.

Firstly, the benchmark analysis implies that SMEs are more likely to outsource parts of the conformity assessment to a third party, either due to limited in-house resources or because of risk aversion that sees them attribute high value to the credibility of an independent assessment by experts.

In the case of the EU-type examination, where all companies must draw up technical documentation to be reviewed by a notified body, SMEs could benefit from being more familiar with this kind of procedure. However, as many SMEs do not have in-house facilities, the testing costs can be significantly higher (in absolute terms but also per unit sold). From the benchmark analysis, testing costs range between EUR 100 and EUR 1,000,000, with a computed average of EUR 10,000. However, experts indicated that this number might be too low and that testing costs would be higher for all enterprises, as no common procedures for testing AI systems and products are in place. All companies - not only SMEs - will bear high additional testing costs in the initial years after the regulation comes into being.

One representative from a start-up association highlighted that additional costs will also arise for consulting legal expertise, as SMEs more often call on external legal consultancy because of a lack of in-house capacity. As legal fees and provisions vary significantly according to the nature of the products and then diminish substantially once the company has adapted to the regulation, costs for legal consultancy were not included in the calculation of the compliance costs. In the long-term, however, the difference between SMEs and large corporations in the cost impact proportionate to their investment is expected to be very small, as increased competition will eventually push costs down.

In the case of full QA, SMEs are less likely to have a QMS in place already and would therefore have to bear high one-off costs to invest in a system to comply with the regulation. The likely significant burden is evident in Mr Cobbaert's estimation of between EUR 80,000 and EUR 160,000 for an organization with 100 employees (the range reflecting the complexity of the organization and the costs for additional external consultants). That one-off spend may deter new entrants from developing high-risk AI systems, but have less impact on existing companies that have already made their investment decision. The investment incentive for existing companies to develop new AI systems should be intact once they have invested in setting up a new QMS. Although the AI regulation may deter some firms from entering the EU market - similar to the impact of the GDPR on certain foreign SMEs - the negative impact brought by the regulation should not be overstated.

It is expected that there will be a larger increase in costs for SMEs in the initial years after the regulation enters into force, as they will need time to familiarise themselves with the requirements and to build up the necessary infrastructure. Nevertheless, costs for SMEs could be significantly reduced by sharing systems (e.g. for testing or legal advice). In the workshop and interviews for this study, representatives of SMEs highlighted that such sharing platforms should be encouraged to help SMEs to navigate the regulation. Some

participants suggested that the Commission or national governments should provide administrative and technical assistance to SMEs and notified bodies at the beginning of the implementation. A centralised learning platform for all players on the market could facilitate responses to the regulation and more organic development of the AI sector.

There are two types of SMEs. The first is AI developers, who are required to ensure compliance and pay for conformity tests of their inventions. The costs will be shared with downstream buyers of the systems, although the developers are supposed to absorb liability and related costs during the deployment phase. This type of company includes deployers who buy customised AI systems, which should be considered co-developers and directly involved in the compliance and conformity process. The second type of SME is AI deployers who purchase certified standardised AI systems and who are thus not supposed to directly pay for compliance and conformity tests of the systems. The costs are shared by various players in the market, with sharing being fair in a sufficiently competitive market. If only high-risk AI systems are subject to the new regulation, the impact on the survival of SMEs will likely be moderate.

When estimating the additional costs for SMEs resulting from the AI legislation, the costs for SMEs in complying with the GDPR are taken as a reference value. According to the SIA Partners Report, GDPR may incur additional costs of GBP 300-GBP450 (EUR 330 – EUR 495) per employee (SIA Partners, 2017). For example, an average SME of 100 employees is estimated to pay EUR 33,000-EUR 49,500 for GDPR compliance. The AI regulation may bring in additional costs but would not make investments substantially unprofitable.

Some other sources of GDPR compliance cost estimates are provided in Table 43 for comparison.

Table 42 - GDPR compliance cost estimates for SMEs (other sources)

Source	Cost estimate
GDPR Small Business Survey (2019) https://gdpr.eu/2019-small-business-survey/	• 51% spent between EUR 1,000 and EUR 50,000 • 18% spent more than EUR 50,000 (up to EUR 1 million)
Datagrail report (2020) https://datagrail.io/downloads/GDPR-CCPA-cost-report.pdf	• Average organisation spent 2,000-4,000 hours in meetings alone to prepare • 74% of SMEs spent more than USD 100,000 • 34% of large enterprises spent more than EUR 1 million
MicroWarehouse Survey (2018) (Hoare, 2018)	• More than four in 10 larger firms in Dublin spent upwards of EUR 20,000 to get ready for the GDPR, while nine out of 10 SMEs spent EUR 5,000
Estimated cost for a small software company (2018) (Fruchte, 2018)	• USD 21,700 for an SME with very strict privacy standards in place and already compliant with existing standards such as Privacy Shield (ongoing costs not included)
Christensen et al. (2013), The Impact of Data Protection Regulation in the EU.	• Estimate that the average SME in the EU can expect its annual cost to increase by between approximately EUR 3,000 and EUR 7,200, depending on the industry (16-40% of current annual SME IT budgets)
SIA Partners (2017) https://sia-partners.co.uk/preparing-gdpr-need-15m-300-450-per-employee-average-implement-gdpr/	• GDPR implementation costs GBP 300-GBP 450 per employee

For comparison with other legislation, the benchmark used here is also relevant for SMEs. The second benchmarking estimation (see Table 36) was derived from the case studies in the Evaluation of Internal Market Legislation for Industrial Products, in which some products represent markets in which SMEs are dominant (e.g. lifts and air conditioners). The cost of conformity assessment for lifts was between EUR 25,200 and EUR 31,000, while for air conditioners it represented EUR 234,702 per year for multiple products. These case studies highlight that SMEs often have to bear higher costs than large companies because of a lower number of units sold, yet they are often more accustomed to using third parties' services and thus some of the additional cost can actually be considered BAU.

9. Assessing the costs of compliance with the forthcoming AI regulation: challenges and limitations

AI is a generic term that encompasses a very diverse set of techniques, including different paradigms (symbolic, statistical, sub-symbolic), methods (logic-based, problem-based, probabilistic, machine learning, embodied intelligence, search and optimisation), and problem domains (perception, reasoning, knowledge, planning, and communication). The multi-purpose use of AI and its ubiquity in many sectors and across a range of physical and digital products challenge not only the regulatory scrutiny of such systems, but also the assessment of compliance and resulting costs. The section below lists the methodological and empirical challenges encountered in the cost estimation of compliance with the AI regulation for Europe. The cost estimations are non-exhaustive, as it is impossible to consider all AI applications and all scenarios here. Equally, the uncertainty around the various regulatory requirements and details of the regulatory text itself significantly limits the design of reliable and grounded cost estimates. To reiterate, the five regulatory requirements put forward by the European Commission in its White Paper are: training data, record-keeping, provision of information, human oversight, accuracy and robustness.

a. Diverse stakeholders engaged with AI systems

AI systems are auxiliary in nature, meaning that there is a range of combinations of AI systems with other hardware and software products. For instance, AI can be a natural language processing system used for an ordinary chatbot giving predefined answers to questions. At the same time, this natural language processing system can be implemented into a sentiment analysis tool that analyses the results of the questions and answers provided by the chatbot. AI systems may be developed as stand-alone products or can likewise be retrained to satisfy another task in a very different context. A distinction is made here between self-developed (in-house) and externally acquired (third-party) AI systems: most likely, AI services will be implemented to improve existing production processes and be added to existing products, software or manufacturing processes. Most of the product manufacturers might not develop their proprietary AI systems but instead purchase AI systems from a third party, such as a software engineering company (developer). This purchased AI system is then added or integrated into the existing product or process by the company (deployer), with or without additional training of the AI system. In addition, acquired third-party AI systems will be retrained with different datasets prior to market release. Larger enterprises and multinational companies are more often developers and deployers of AI systems at the same time by establishing in-house AI development teams. Overall, diverse roles by different stakeholders make it difficult to establish clear responsibilities with regard to compliance activities. For the purposes of the cost estimation study, three (non-exhaustive) scenarios were considered:

Case 1. A company develops an AI system and seeks certification.

Case 2. A company purchases an AI system from an upstream firm without changes to the code or the training datasets. The firm embeds this third-party AI system into a product and seeks to certify this product.

Case 3. A company purchases an AI system from an upstream firm and changes the code and/or the training datasets. The firm embeds this modified third-party AI system into a product and seeks to certify this product.

The difficulties encountered in assessing the costs for the three scenarios are discussed below. As the proposed regulation relies heavily on the use and keeping of data in the training process of AI systems, the compliance and conformity assessment cost estimations are closely dependent on the volume and source of the data.

Case 1, a company designs, develops and uses an AI system in-house: This company represents both the developer and deployer, being responsible for fulfilling all obligations stemming from the regulatory requirements. This includes appropriate training of data, record-keeping, information provision procedures, robustness and accuracy checks, and human oversight measures.

Case 2, a company (deployer) acquires an AI system from a third party (developer): The AI system is not further changed and thus the training data required to develop the AI system is owned by the developer only. The responsibility to keep appropriate records of the training data thus lies with the developer, who should then seek certification of the AI system. However, the deployer acquiring the third-party AI system is also responsible for demonstrating regulatory compliance. It remains unclear whose responsibility it is to demonstrate regulatory compliance: in principle, both parties should share the administrative burden of the conformity assessment procedure through legal and contractual arrangements. However, it would be less costly if the developer sought certification for its AI system independently of the deployer because the AI system might also be sold to other deployers. To conclude, a point to be clarified is the flexibility and responsibility of certifications between AI developers and deployers.

Case 3, a company (deployer) acquires an AI system from a third-party (developer) and further retrains the AI system with a separate dataset in-house: Both the deployer and the developer are involved in the training data process for the AI system. In this case, it is unclear how a notified body would assess compliance with the training data requirements if both developer and deployer used different datasets. It is assumed that it would be insufficient for the notified body to verify only the third-party acquired AI system without verifying the retrained AI with the dataset of the deployer. The notified body is thus likely to audit the training dataset from both developer and deployer before issuing a certificate.

Alternatively, regulatory compliance of an AI system could be assessed in two separate conformity assessments. Assessing two individual AI systems including the data is feasible, although not ideal. The least costly approach would be to ask the upstream AI-component developer and the downstream AI-embedded deployer to seek separate certifications that audit their corresponding training data. The developer provides the certification for the deployer to facilitate its own conformity assessment process. In cases where the deployer provides most of the data, the developer may not see an advantage in seeking certification, so the deployer should consider owning and keeping records of the data provided by the developer.

An AI system may rely on many open-source inputs and pre-trained systems. Developers may encounter difficulties in providing information about the training data and other technical information on the AI system.

b. Expected conformity assessment procedure performed by notified bodies

The expert interviews and the high-level workshop suggest that notified bodies seldom if ever perform type examination of products containing software under the MDR. This is because of the complexity of certifying software. Apart from the lack of expertise, the main

obstacle is to exhaustively test the software. Notified bodies might bear significant legal consequences if a harmful defect was not detected during the assessment process. A type examination for certification of software - and thus AI systems – does not seem feasible given the current expertise level of notified bodies and the existing technology level to test software.

It is unclear whether the AI regulation would require actual auditing of training data and record-keeping. Stakeholders suggested that testing the outputs or the performance of an AI product would already be sufficient for inferring compliance with the requirement on training data because the outputs are essentially what matters to AI end users. Another issue is that a representative training dataset may not necessarily lead to unbiased or non-discriminatory decisions.

A related question concerns the expected testing to be performed by notified bodies. Currently, there are no widely acceptable standards for AI systems. An AI regulation may have to be accompanied by a governance standard, a risk management standard and a robustness and accuracy standard. With clear guidance on assessment and better legal protection for both applicants and notified bodies, such bodies may be more willing to conduct conformity assessment of AI products (AI components are basically software).

c. One-off vs concurrent costs

With neither the requirements or the testing procedure/standards yet well-defined, it is almost futile to estimate the one-off cost of compliance procedures with the regulatory requirements for the EU economy. Both industry stakeholders and notified bodies would need to adjust their procedures to comply with the new regulation. For AI providers, one-off costs may include staff training, legal consultancy fees, and any machinery or equipment needed for compliance. Notified bodies may also have to upgrade their expertise and equipment, and some new bodies may be established specialising in certifying AI products. The report focuses mainly on concurrent costs, assuming that the industry has already adjusted to the new norm.

A substantial volume of costs would stem from setting up internal (QMS) for companies, particularly SMEs. As the proposed regulation may reach all industries, many firms in lightly regulated industries might need to invest in a substantial one-off cost for market entry, effectively setting up an entry barrier and dampening market competition.

Stakeholders and notified bodies believe these estimates to be notional, with the reality becoming evident only in the longer term. They stress the cost of staff training in the initial stages of regulation, with learning-by-doing lowering the cost over time. On the other hand, lack of competition among notified bodies, together with a huge demand for assessments from applicants in the early stages, would push up costs (prices).

d. Compatibility with existing conformity assessment

The MDR was adopted in May 2017 and will replace the current Medical Device Directive (MDD) from 26 May 2021. According to the MDR, classification as a medical device depends on the **manufacturer** specifying the intended use. More relevant in the AI regulation context is the fact that almost all software as a medical device (SaMD) will be moved to higher risk classes under rule 11 of the MDR. More specifically, all software used ‘to take decisions with diagnosis or therapeutic purposes’ will be class IIa at a minimum (Chaper III, Rule 11, MDR). Software can be classified as higher risk if it has the potential to cause serious deterioration of a person’s health (class IIb) or death/irreversible deterioration (class III) (Decomplix, 2019).

It is far from evident how the AI regulation would be imposed on products that are subject to existing regulations and conformity assessment procedures. In principle, the two

certification processes could be done separately, but in practice they would likely be conducted simultaneously. Any savings stemming from merging two certification processes, or indeed any additional costs incurred for assessing the AI component, remain unknown.

Another question is the association between procedures of regulations. In theory, the applicant can apply for certification through Annex IX (Full QMS) under MDR while applying for an EU-type examination of the AI component. Again, the question of whether the two procedures are compatible or the possibility of a harmonised certification process for an AI-embedded product remains unanswered.

For toy safety certification, producers or importers are required to pass an EU-type examination before their products can be sold in the EU market (Article 20 of Directive 2009/48/EC). The same applicant could, in principle, apply an EU-type examination for an AI-component.

Feedback from medical device market stakeholders points to one particular difficulty with a type examination procedure - capability and reluctance to conduct a type examination of a product containing software. Notified bodies lack expertise and exhaustive testing of software is very costly and difficult, if not impossible, given the current technology level. Notified bodies are also wary of the potential legal consequences of any undetected problems with the product. Devices with software usually go through the assessment procedure based on QMS and review of technical documentation. It raises the concern that type examinations may not be a realistic procedure for certification. Even if the regulation shifts the burden of testing of the product from notified bodies to applicants, the applicant should provide extensive information on the in-house testing as proof to the notified body. This ideally less costly procedure for SMEs may nevertheless be difficult to pursue.

If no notified body is willing to conduct a required EU-type examination of an AI system, the toy producer could only go through the full QA procedure to obtain certification. A challenge for authorities and notified bodies is to equip themselves with internal competence to critically evaluate AI technologies, even if some assessments could be conducted by external experts. Some EU-wide preparation for AI providers and notified bodies would help lower the overall costs.

e. Legal costs

The study team made the methodological choice to exclude the costs of external legal advice[88] and consultancy fees from the cost estimates. This choice is grounded in the observation that these cost items are largely influenced by (i) the size of a company and the availability of in-house expertise, (ii) the preference of each individual company, and (iii) the complexity and intrusiveness of the regulatory requirements in the proposed regulation. The digital survey results, as well as the follow-up interviews after the high-level workshop, revealed that industry members consider legal and consultancy fees an important element of compliance costs. Industry stakeholders recommended that instead of calculating a lump-sum estimate for legal fees, these cost items should be factored into the compliance cost estimates under each regulatory requirement separately.

In general, many companies indicated that (external) legal advice will be necessary for all compliance activities. They argue that the regulatory gaps, lack of best practice in the industry, and the complexity of regulatory definitions necessitate hiring legal experts and consultants to interpret and understand the details of the regulation. All stakeholders

[88] For the purposes of this report, legal fees are understood as specialised expert advice on compliance. As such, they exclude potential litigation fees for breaches of regulatory requirements.

generally agreed that SMEs would be disproportionately affected by these costs, considering the lack of available in-house expertise.

The results of the desk research and the survey results indicated a broad scale for legal expert fees, ranging from EUR 250 to EUR 1,600 per hour. Lawyers generally charge hourly rates, whereas some consultancies charge on a per-project basis (ranging from the low thousands to hundreds of thousands of euro).

Calculating the number of hours of legal advice a company would need to ensure legal compliance for a single AI product is not yet possible, nor is there any market price for AI compliance projects carried out by consultancies. Incorporating these cost items under the Standard Cost Model thus goes beyond the scope of this report.

It is possible, however, to approximate the type of legal advice companies will need, based on the survey responses. Respondents most often mentioned legal fees as an important cost item under three regulatory requirements: (i) training data, (ii) keeping data and records, and (iii) information provision obligations.

Industry members frequently noted that specialised AI/privacy legal expertise is necessary to ensure that compliance activities under the AI regulation do not jeopardise compliance with other regulations, in particular, the GDPR and other privacy laws.

Under the ***training data*** requirement, respondents raised the issue of using non-personalised datasets for training AI algorithms. They argued that it requires significant additional resources to bring the additional data used to achieve compliance with the data quality standards of the AI regulation in line with GDPR standards. Respondents also indicated that the cost of compliance with the training data requirement depends heavily on the availability of first-party data in-house. Where an organisation primarily procures third-party data, legal advice becomes necessary to ensure that the company has acquired and processed data lawfully.

Under the **keeping of records and data** requirement, respondents argued that external legal advice is necessary to assess the legal implications of preserving documentations and datasets in light of data minimisation requirements under the GDPR. Even though such documentation would be integrated into internal GDPR compliance processes, there is an added cost of legal advice.

Under the **information provision** requirement, respondents observed that if this obligation confers the consumer right to request additional information from AI providers, external legal advice will be necessary to comply with those requests. In addition, companies would need specialised legal advice in order to avoid abuse or fraudulent use of such information (e.g. exposing trade secrets or carrying out attacks on an AI product's cybersecurity system).

10. Conclusion

This report estimates that total compliance cost of the proposed regulation on AI systems is roughly 17% of total AI investment cost. Projection to the population shows that compliance with the proposed regulation may cost the EU economy EUR 131 million – EUR 345 million in 2022, and the global economy EUR 593 million – EUR 1.569 billion under a high-risk (10% coverage) only AI regulation. Conformity assessment would entail another 13.5% of AI investment cost, while setting up a QMS may have an upfront cost of up to EUR 330,000 per firm.

These estimates depend on the evolution of the AI market and the definition of high-risk applications by the proposed regulation. This report also looked at some challenges facing the regulatory authorities.

REFERENCES

Access Now (2018). Human Rights in the Age of Artificial Intelligence. Available at: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

ACOLA (2019). The effective and ethical development of artificial intelligence: An opportunity to improve our wellbeing. Available at: https://acola.org/wp-content/uploads/2019/07/hs4_artificial-intelligence-report.pdf

Adadi, A. and Berrada, M. (2018). 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)' Available at: https://ieeexplore.ieee.org/document/8466590

AI HLEG (2019). Ethics guidelines for trustworthy AI. EU High Level Expert Group on AI. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

AI HLEG (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) self-assessment. EU High Level Expert Group on AI. Available at: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

AI Industry Alliance (2019). Joint Pledge on Artificial Intelligence Industry Self-Discipline. China. Available (in Chinese) at: https://mp.weixin.qq.com/s/x7HTx4AR6oNBWwWxUpnSuQ

AI Law (n.d.). AI and discrimination. Available at: https://ai-lawhub.com/ai-and-discrimination/

AI Now (2019). '2019 Report'. Available at: https://ainowinstitute.org/AI_Now_2019_Report.pdf

AI Now (2020). *Regulating Biometrics.* Available at: https://ainowinstitute.org/regulatingbiometrics.pdf

AI Prescience (2019). '5 ways Starbucks uses data to gain competitive edge'. *AI Prescience*. 12 July 2019. Available at: https://www.aiprescience.com/how-starbucks-uses-data-and-ai/

Azizi, S. and Yektansani, K. (2020). 'Artificial Intelligence and Predicting Illegal Immigration to the USA.' International Migration, 58(5), 183–193. Available at: https://doi.org/10.1111/imig.12695

Alfeld (2016). 'Data Poisoning Attacks against Autoregressive Models'. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/10237

AlgorithmWatch (2019). 'Personal Scoring in the EU: Not quite Black Mirror yet, at least if you're rich'. Available at: https://algorithmwatch.org/en/personal-scoring-in-the-eu-not-quite-black-mirror-yet-at-least-if-youre-rich/

AlgorithmWatch (2020). 'ADM systems in the covid-19 pandemic: A European Perspective'. Available at: https://algorithmwatch.org/en/project/automating-society-2020-covid19/

AlgorithmWatch (n.d.). AI ethics guidelines global inventory. Available at: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

Allen, R. and Masters, D. (2020). *Regulating for AI. A new role for equality bodies*. Equinet: European Network of Equality Bodies. Available at: https://equineteurope.org/wp-content/uploads/2020/06/ai_report_digital.pdf

Allied Market Research (2018). Artificial Intelligence (AI): Market Outlook. Available at: https://www.alliedmarketresearch.com/artificial-intelligence-market

Amodei, D. et al. (2016). 'Concrete Problems in AI Safety'. Available at: https://arxiv.org/abs/1606.06565

Angwin, J., Mattu, S. and Larson, J. (2015). 'The tiger mom tax: Asians are nearly twice as likely to get a higher price from Princeton Review'. *ProPublica*. Available at: https://www.ProPublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016). 'Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks.' *ProPublica*. Available at: www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencingwww.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arnold, M. et al. (2019). 'FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity'. Available at: https://arxiv.org/abs/1808.07261

Australian Government (2019). 'Artificial Intelligence: Australia's Ethics Framework'. Available at: https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/

Babuta, A., Oswald, M. and Rinik, C. (2018). 'Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges'. Available at: https://rusi.org/publication/whitehall-reports/machine-learning-algorithms-and-police-decision-making-legal-ethical

Bahara, H., Kranenberg, A. and Tokmetzis, D. (2019). 'Hoe YouTube rechtse radicalisering in de hand werkt.' *De Volkskrant*.

Barber, G. (2019a). 'Artificial intelligence confronts a "reproducibility" crisis.' *Wired.* Available at: https://www.wired.com/story/artificial-intelligence-confronts-reproducibility-crisis/

Barber, G. (2019b). 'San Francisco bans agency use of facial-recognition tech.' *Wired.* Available at: https://www.wired.com/story/san-francisco-bans-use-facial-recognition-tech/

Barocas, S. and Nissenbaum, H. (2014). 'Big Data's End Run around Anonymity and Consent'. Available at: https://www.cambridge.org/core/books/privacy-big-data-and-the-public-good/big-datas-end-run-around-anonymity-and-consent/0BAA038A4550C729DAA24DFC7D69946C

Barocas, S. and Selbst, A.D. (2016). 'Big Data's disparate impact'. *California Law Review 104(3),* p.671-732.

Beduschi, A. (2020). 'International migration management in the age of artificial intelligence.' *Migration Studies*, mnaa003. Available at: https://doi.org/10.1093/migration/mnaa003https://doi.org/10.1093/migration/mnaa003

Beglinger, C. (2019). 'Note: A Broken Theory: The Malfunction Theory of Strict Products Liability and the Need for a New Doctrine in the Field of Surgical Robotics'. Available at:

https://minnesotalawreview.org/article/note-a-broken-theory-the-malfunction-theory-of-strict-products-liability-and-the-need-for-a-new-doctrine-in-the-field-of-surgical-robotics/

Biocca, F. (1997). 'The Cyborg's Dilemma: Progressive Embodiment in Virtual Environments'. Available at: https://academic.oup.com/jcmc/article/3/2/JCMC324/4080399?login=true

Bodkin et al. (2020). 'Visuomotor therapy modulates corticospinal excitability in patients following anterior cruciate ligament reconstruction: A randomized crossover trial'. Available at: https://www.researchgate.net/publication/347083728_Visuomotor_therapy_modulates_corticospinal_excitability_in_patients_following_anterior_cruciate_ligament_reconstruction_A_randomized_crossover_trial

Bourne, C. D. (2019). 'AI cheerleaders: Public relations, neoliberalism and artificial intelligence.'. Available at: https://research.gold.ac.uk/id/eprint/25962/1/AI-Neoliberalism-PR-Jan2019%20(1).pdf

Bradley et al. (2020). 'National Artificial Intelligence Strategies and Human Rights: A Review'. Available at: https://www.gp-digital.org/wp-content/uploads/2020/04/National-Artifical-Intelligence-Strategies-and-Human-Rights%E2%80%94A-Review_.pdf

Brantingham, P.J. (2018). 'The logic of data bias and its impact on place-based predictive policing.' *Ohio State Journal of Criminal Law* (15), 473-486. Available at: https://kb.osu.edu/bitstream/handle/1811/85819/1/OSJCL_V15N2_473.pdf

Brown, T.B., Mann, B., Ryder, N. and Subbiah, M. (2020). 'Language models are few-shot learners.' *ArXiv.* Available at: https://arxiv.org/pdf/2005.14165.pdf

Bryson, J. (2019). 'A smart bureaucrat's guide to AI regulation.' Blog. 16 January 2019. Available at: https://joanna-bryson.blogspot.com/2019/01/a-smart-bureaucrats-guide-to-ai.html

Bullock et al. (2020). 'Mapping the landscape of Artificial Intelligence applications against COVID-19'. Available at: https://www.jair.org/index.php/jair/article/view/12162

Buolamwini, J. and Gebru, T. (2018). 'Gender Shades: Intersectional accuracy disparities in commercial gender classification'. Proceedings of Machine Learning Research 81, p.1-15.

BusinessWire (2019). Global $52Bn biometric authentication & identification market, 2023: focus on modality, motility, application and technology - ResearchAndMarkets.com. Available at: https://www.businesswire.com/news/home/20190410005486/en/Global-52Bn-Biometric-Authentication-Identification-Market-2023

Calders, T., Kamiran F., Pechenizkiy, M. (2009). 'Building Classifiers with Independency Constraints'. Available at: https://ieeexplore.ieee.org/abstract/document/5360534

Caliskan, A., Bryson, J. and Narayanan, A. (2017). 'Semantics derived automatically from language corpora contain human-like biases'. *Science*, 356, 183-186.

Calmon F.P. et al. (2017). 'Optimized Data Pre-Processing for Discrimination Prevention'. Available at: https://arxiv.org/abs/1704.03354

Calo R. (2013). 'Digital Market Manipulation'. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2309703

Campolo, A., Sanfillipo, M., Whittaker, M. and Crawford, K. (2017). *AI Now 2017 report*. Available at: https://ainowinstitute.org/AI_Now_2017_Report.pdf

Canetti, R. et al. (2019). 'From Soft Classifiers to Hard Decisions: How fair can we be?'. Available at: https://dl.acm.org/doi/abs/10.1145/3287560.3287561

Carlini, N. et al. (2019). 'On Evaluating Adversarial Robustness'. Available at: https://arxiv.org/abs/1902.06705

CCDCOE (n.d.). 'CJEU declares general data retention unlawful in Tele2 Sverige.NATO Cooperative Cyber Defence Centre of Excellence.Tallinn. Available at: https://ccdcoe.org/incyder-articles/cjeu-declares-general-data-retention-unlawful-in-tele2-sverige/

Centre for Humane Technology (n.d.). App Ratings. Available at: https://www.humanetech.com/app-ratings

Cerulus, L. (2019). 'How Ukraine became a test bed for cyberweaponry.' *Politico.* Available at: https://www.politico.eu/article/ukraine-cyber-war-frontline-russia-malware-attacks/

Chowdhury, A.R. et al. (2020). 'Data Drift and Machine Learning Model Sustinability'. Available at: https://www.analyticsinsight.net/data-drift-and-machine-learning-model-sustainability/

Christensen et al. (2013). 'The impact of data protection regulation in the EU'. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.657.138&rep=rep1&type=pdf

Christiano, P. (2016). 'Semi-supervised reinforcement learning'. Available at: https://medium.com/ai-control/semi-supervised-reinforcement-learning-cf7d5375197f

Cihon, P. (2019). 'Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development'. Available at: http://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf

Clifford, D. (2019). The legal limits to the monetisation of online emotions. Available at: https://lirias.kuleuven.be/2807964

Cognilytica (n.d.). Classification of the AI vendor ecosystem. Available at: https://www.cognilytica.com/2019/01/16/cognilyticas-classification-of-the-ai-vendor-ecosystem-overview-and-bottom-3-layers/

Cohen, J. (2018). 'Exploring Echo-Systems: How Algorithms Shape Immersive Media Environments.' *Journal of Media Literacy Education,* 10(2), 139-151. Available at: https://eric.ed.gov/?id=EJ1198674

Corea, F. (2019). 'AI Knowledge Map: How To Classify AI Technologies'. Available at: https://www.forbes.com/sites/cognitiveworld/2018/08/22/ai-knowledge-map-how-to-classify-ai-technologies/?sh=441a73857773

Council of Europe (2017). *Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications*. Committee of Experts on Internet Intermediaries. Available at: https://rm.coe.int/study-hrdimension-of-automated-data-processing-incl-algorithms/168075b94a

Council of Europe (2018). *Algorithms and human rights*. Available at: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

Council of Europe (2019a). *Ethnic profiling: a persisting practice in Europe*. Council of Europe Commissioner for Human Rights. Available at: https://www.coe.int/en/web/commissioner/-/ethnic-profiling-a-persisting-practice-in-europe

Council of Europe (2019b). *Unboxing artificial intelligence: 10 steps to protect human rights*. Council of Europe Commissioner for Human Rights. Available at: https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64

Crumpler, W. (2020). 'How accurate are facial recognition systems – and why does it matter?' *Technology Policy Blog*. Centre for Strategic and International Studies (CSIS). Available at: https://www.csis.org/blogs/technology-policy-blog/how-accurate-are-facial-recognition-systems-—-and-why-does-it-matter

Dafoe, A. and Zwetsloot, R. (2019). 'Thinking About Risks From AI: Accidents, Misuse and Structure'. Available at: https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure

Datagrail (2020). The age of privacy: the costs of continuous compliance. Available at: https://datagrail.io/downloads/GDPR-CCPA-cost-report.pdf

Datenethikkommission (2019). Data Ethics Commission of the German Federal Government (Datenethikkommission), Opinion (English version), Berlin, December 2019.

Datta, A., Tschantz, C. and Datta, A. (2015). 'Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination'. *Proceedings on Privacy Enhancing Technologies* 2015(1).

Dawson, M., Burrell, D.N., Rahim, E., and Brewster, S. (2010). 'Integrating Software Assurance into the Software Development Life Cycle (SDLC)'. Available at: https://www.researchgate.net/publication/255965523_Integrating_Software_Assurance_into_the_Software_Development_Life_Cycle_SDLC

Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J. and Hajkowicz, S. (2019). 'Artificial Intelligence: Australia's Ethics Framework'. *Data61 CSIRO*, Australia. Available at: https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdfhttps://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf

Decomplix (2019). 10 facts you need to know about the MDR as a medical software manufacturer. Available at: https://decomplix.com/medical-software-mdr/

Deepmind (n.d.). Ethics and society principles. Available at: https://deepmind.com/applied/deepmind-ethics-society/principles/

De Souza, W.G. et al. (2019). 'How and where is artificial intelligence in the public sector going? A literature review and research agenda'. Available at: https://www.researchgate.net/publication/334685240_How_and_where_is_artificial_intelligence_in_the_public_sector_going_A_literature_review_and_research_agenda

D'Ignazio, C. and Klein L.F. (2019). 'Data Feminism'. Available at: https://mitpress.mit.edu/books/data-feminism

Dupré, D. et al. (2020). 'A performance comparison of eight commercially available automatic classifiers for facial affect recognition'. Available at: https://pubmed.ncbi.nlm.nih.gov/32330178/

ECHA (2019). European Chemicals Agency final annual accounts for financial year 2019. Available at: https://echa.europa.eu/documents/10162/13611/echa_annual_accounts_2019_en.pdf/1e6f1f63-21a7-1db2-11e5-9c60cc8ac44e

ECHA (2020). European Chemicals Agency Budget 2020. Available at: https://echa.europa.eu/documents/10162/28676836/FINAL_MB_58_2019_Budget_2020_MB56.pdf/5b13753a-9352-c2d3-e7a9-22721e15c981

EDPS (2017). EDPS Opinion on the proposal for a Regulation on ECRIS-TCN. Opinion 11.2017. European Commission. Brussels. Available at: https://edps.europa.eu/sites/edp/files/publication/2017_0542_draft_opinion_ecris_tcn_revab_en.pdf

EDPS (2020). Artificial intelligence, data and our values – on the path to the EU's digital future. European Data Protection Supervisor. Available at: https://edps.europa.eu/press-publications/press-news/blog/artificial-intelligence-data-and-our-values-path-eus-digital_en

EDRi (n.d.). List of articles and documents on the issue of facial and biometric recognition. Available at: https://edri.org/our-work/facial-recognition-document-pool/

Eireiner, V. (2020). 'Imminent dystopia_ Media coverage of algorithmic surveillance at Berlin-Südkreuz'. Avaiable at: https://www.researchgate.net/publication/340413439_Imminent_dystopia_Media_coverage_of_algorithmic_surveillance_at_Berlin-Sudkreuz

Ellahham S. et al. (2019). 'Application of Artificial Intelligence in the Health Care Safety Context: Opportunities and Challenges'. Available at: https://www.researchgate.net/publication/336262852_Application_of_Artificial_Intelligence_in_the_Health_Care_Safety_Context_Opportunities_and_Challenges

EMA (n.d.) Fees payable to the European Medicines Agency. Available at: https://ec.europa.eu/info/energy-climate-change-environment/standards-tools-and-labels/products-labelling-rules-and-requirements/energy-label-and-ecodesign/product-database_en

Engler, A. (2019). 'For some employment algorithms, disability discrimination by default.' *Brookings*. Available at: https://www.brookings.edu/blog/techtank/2019/10/31/for-some-employment-algorithms-disability-discrimination-by-default/

Engstrom, D.F., Ho, D.E., Sharkey, C.M. and Cuellar, M-F. (2020). Government by Algorithms: Artificial Intelligence in Federal Administrative Agencies, Report submitted to the Administrative Conference of the United States.

ENISA (2019). European Union Agency for Cybersecurity 2019 annual budget. Available at: https://www.enisa.europa.eu/about-enisa/accounting-finance/files/annual-budgets/enisa-2019-annual-budget/view

EPIC (2020). AI and human rights: criminal justice system. Electronic Privacy Information Centre. Available at: https://epic.org/ai/criminal-justice/

EPRS (2020). *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*. European Parliament Research Service, Brussels. Available at: https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU%2820%29641530

EPRS (2020). *Civil liability regime for Artificial Intelligence*. European Parliamentary Research Service.

European Commission (2014a). A vision for the internal market for products. Accompanying the Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee. SWD(2014) 23 final. Part 2/2. Brussels 22.1.2014. Available at: https://eur-lex.europa.eu/resource.html?uri=cellar:6da8f15b-8438-11e3-9b7d-01aa75ed71a1.0001.05/DOC_1&format=PDF

European Commission (2014b). Commission staff working document Part 1: Evaluation of the Internal Market Legislation for Industrial Products. SWD/2014/023 final. Available at: https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:52014SC0023

European Commission (2017). Impact assessment accompanying the document 'Proposal for a Regulation of the European Parliament and of the Council on ENISA, the "EU Cybersecurity Agency"', and repealing Regulation (EU) 526/2013, and on Information and Communication Technology cybersecurity certification ('Cybersecurity Act'). SWD/2017/0500 final - 2017/0225 (COD). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52017SC0500

European Commission (2018). Tackling Online Disinformation. Communication.

European Commission (2019). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Building Trust in Human-Centric Artificial Intelligence. COM(2019) 168 final. Available at: https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

European Commission (2019a). Liability for Artificial Intelligence. Report from the Expert Group on Liability and New Technologies – New Technologies Formation. Available at: https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608

European Commission (2020a). White Paper on Artificial Intelligence – A European approach to excellence and trust. COM(2020) 65 final. Available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission (2020b). *Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics*. Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee. COM(2020) 64 final. Available at: https://ec.europa.eu/info/sites/info/files/report-safety-liability-artificial-intelligence-feb2020_en_1.pdf

European Group on Ethics in Science and New Technologies (2019). EGE sets out constructive ways forward for ethical AI. Available at: https://ec.europa.eu/info/news/ege-sets-out-constructive-ways-forward-ethical-ai-2019-feb-05_en

European Parliament (2019). A governance framework for algorithmic accountability and transparency. Scientific Foresight Unit (STOA) Briefing. European Parliament. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262(ANN1)_EN.pdf

European Parliament (2020). 'Parliament leads the way on first set of EU rules for Artificial Intelligence'. Available at: https://www.europarl.europa.eu/news/en/press-room/20201016IPR89544/parliament-leads-the-way-on-first-set-of-eu-rules-for-artificial-intelligence

European Patent Office (2019). Guidelines for examination in the European Patent Office. Available at: https://www.epo.org/law-practice/legal-texts/guidelines.html

Fanni, R., Steinkogler, V.E., Zampedri, G. and Pierson, J. (2020). 'Active human agency in artificial intelligence mediation.' *GoodTechs '20: Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*. Available at: https://doi.org/10.1145/3411170.3411226

FATML (2016). Principles for accountable algorithms and a social impact statement for algorithms. Available at: https://www.fatml.org/resources/principles-for-accountable-algorithms

Favaretto, M., De Clercq, E. and Elger, B.S. (2019). 'Big Data and discrimination: perils, promises and solutions. A systematic review'. *Journal of Big Data,* 6(12). Available at: https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-019-0177-4.pdf

Feige, I. (2019). 'What is AI safety?'. Available at: https://faculty.ai/blog/what-is-ai-safety/

Feldman Barrett, L., Adolphs, R., Marsella, S., Martinez, A.M. and Pollak, S.D. (2019). 'Emotional expressions reconsidered: challenges to inferring emotion from human facial movements.' *Psychological Science in the Public Interest,* 20(1). Available at: https://www.psychologicalscience.org/publications/emotional-expressions-reconsidered-challenges-to-inferring-emotion-from-human-facial-movements.html

Ferguson, A.G. (2017). 'The Rise of Bid Data Policing'. Available at: https://www.degruyter.com/document/doi/10.18574/9781479854608/html

Ferretti, F. (2017). The never-ending European credit data mess. BEUC: European Consumer Organisation. Available at: https://www.beuc.eu/publications/beuc-x-2017-111_the-never-ending-european-credit-data-mess.pdf

Finck, M. (2020). 'Automated decision-making and administrative law.' In P. Cane et al. (eds), *Oxford Handbook of Comparative Administrative Law*. Oxford: Oxford University Press. Max Planck Institute for Innovation & Competition Research Paper No. 19-10. Available at https://ssrn.com/abstract=3433684

Fjelland, R. (2020). 'Why general artificial intelligence will not be realised'. *Humanit Soc Sci Commun,* 7(10). Available at: https://doi.org/10.1057/s41599-020-0494-4

Fletcher, R. (n.d.). 'The truth behind filter bubbles: Bursting some myths.' *Reuters Institute, University of Oxford.* Available at: https://reutersinstitute.politics.ox.ac.uk/risj-review/truth-behind-filter-bubbles-bursting-some-myths

FRA (2018). *#BigData: Discrimination in data-supported decision-making*. European Union Agency for Fundamental Rights. Available at: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-focus-big-data_en.pdf.

FRA (2019). *Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights*. European Union Agency for Fundamental Rights, Vienna. Available at: https://fra.europa.eu/en/publication/2019/data-quality-and-artificial-intelligence-mitigating-bias-and-error-protect

Fruchte, J. (2018). 'Cost of GDPR compliance for a small software business.' *Medium.* 4 October 2018. Available at: Cost of GDPR Compliance for a Small Software Business | by Expected Behavior | Expected Behavior Blog | Medium

Future of Life Institute (2017). Asilomar AI principles. Available at: https://futureoflife.org/ai-principles/

GDPR.EU (2019). GDPR Small Business Survey. Available at: https://gdpr.eu/2019-small-business-survey/

German Data Ethics Commission (GDEC)(2020). Opinion of the Data Ethics Commission. Available at: https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3

German Federal Government, Statistical Office and National Regulatory Control Council (2018). Guidelines on the identification and presentation of compliance costs in legislative proposals by the (German) Federal government. Available in English at: https://www.destatis.de/EN/Themes/Government/Bureaucracy-Costs/Download/ComplianceCostsGuidelines.pdf?__blob=publicationFile

Giannopoulou, A. (2020). 'Algorithmic systems: the consent is in the detail?' *Internet Policy Review* 9(1). Available at: https://policyreview.info/articles/analysis/algorithmic-systems-consent-detail

Gibney, E. (2019). 'This AI researcher is trying to ward off an AI reproducibility crisis.' *Nature.* 19 December 2019. Available at: https://www.nature.com/articles/d41586-019-03895-5

Gibson D. (2020). 2019 Artificial intelligence and automated systems annual legal review. Available at: https://www.gibsondunn.com/wp-content/uploads/2020/02/2019-artificial-intelligence-and-automated-systems-annual-legal-review.pdf

Goethe, T.S. (2019). 'Bigotry encoded: Racial bias in technology.' *Reporter.* Available at: https://reporter.rit.edu/tech/bigotry-encoded-racial-bias-technology

Gonzalez F. (2020). 'Artificial Intelligence and Law Enforcement - Impact on Fundamental Rights'. Available at: https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2020)656295

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). 'Generative Adversarial Nets'. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc.

Government of Canada (2018). Canada-France Statement on Artificial Intelligence. Available at: https://www.international.gc.ca/world-monde/international_relations-relations_internationales/europe/2018-06-07-france_ai-ia_france.aspx?lang=eng

Graber, C. (2018). 'Artificial Intelligence, Affordances and Fundamental Rights'. Available at:

https://www.researchgate.net/publication/330423326_Artificial_Intelligence_Affordances_and_Fundamental_Rights

Grand View Research (2020). Artificial intelligence market size worth $733.7 billion by 2027. Press release. Available at: https://www.grandviewresearch.com/press-release/global-artificial-intelligence-ai-market

Gstrein, O. J., Bunnik, A. and Zwitter, A. (2019). 'Ethical, legal and social challenges of predictive policing.' *Católica Law Review*, 3(3), 77-98.

Gu et al. (2017). 'BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain'. Available at: https://arxiv.org/abs/1708.06733

Guariglia, M. and Quintin, C. (2020). 'Thermal imaging cameras are still dangerous dragnet surveillance cameras.' *Electronic Frontier Foundation*. Available at: https://www.eff.org/deeplinks/2020/04/thermal-imaging-cameras-are-still-dangerous-dragnet-surveillance-cameras

Guidotti et al. (2018). 'A Survey of Methods for Explaining Black Box Models'. Available at: https://dl.acm.org/doi/10.1145/3236009

Gundersen, O.E., Gil, Y. and Aha, D.W. (2018). 'On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications.' *AI Magazine* 39(3). Available at: https://www.isi.edu/~gil/papers/gundersen-etal-aimagazine18.pdf

Habli, I., Lawton, T. and Porter, Z. (2020). 'Artificial intelligence in healthcare: accountability and safety.' *World Health Organization Bulletin*. Available at: https://www.who.int/bulletin/volumes/98/4/19-237487/en/

Harris, M. (2018). 'An eye-scanning lie detector is forging a dystopian future.' *Wired.* 12 April 2018. Available at: https://www.wired.com/story/eye-scanning-lie-detector-polygraph-forging-a-dystopian-future/

Heitmüller, U. (2019). 'Missing Link: Predictive Policing – die Kunst, Verbrechen vorherzusagen.' *Heise online.* Available at: https://www.heise.de/newsticker/meldung/Missing-Link-Predictive-Policing-die-Kunst-Verbrechen-vorherzusagen-4425204.html?seite=all

Helberger, N. and Trilling, D. (2017). 'Facebook is as news editor: the real issues to be concerned about'. Available at: http://eprints.lse.ac.uk/81314/1/Facebook%20is%20a%20news%20editor_%20the%20real%20issues%20to%20be%20concerned%20about%20_%20LSE%20Media%20Policy%20Project.pdf

Hoare, P. (2018). 'MicroWarehouse Survey'. *Irish Examiner,* 26 June 2018. Available at: https://www.irishexaminer.com/business/arid-30851103.html

Hoffmann-Riehm, W. (2020). 'Artificial Intelligence as a Challenge for Law and Regulation'. Available at: https://link.springer.com/chapter/10.1007/978-3-030-32361-5_1

Hollnagel E. (2014). 'From Safety-I to Safety-II: A White Paper'. Available at: https://www.england.nhs.uk/signuptosafety/wp-content/uploads/sites/16/2015/10/safety-1-safety-2-whte-papr.pdf

Hu Y. et al. (2020). 'The challenges of deploying artificial intelligence models in a rapidly evolving environment'. Available at: https://www.nature.com/articles/s42256-020-0185-2

Hunton, A.K. (2020). 'Portland, Oregon First to Ban Private-Sector Use of Facial Recognition Technology.' *Hunton Privacy Blog*. Available at: https://www.huntonprivacyblog.com/2020/09/10/portland-oregon-becomes-first-jurisdiction-in-u-s-to-ban-the-commercial-use-of-facial-recognition-technology/

Hutson, M. (2018). 'Artificial intelligence faces reproducibility crisis.' *Science.* Available at: https://science.sciencemag.org/content/359/6377/725

ICO (n.d.). Explaining decisions made with AI. Information Commissioner's Office UK. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/

ICO (2019). Guidance on AI and data protection. Information Commissioner's Office UK. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/guidance-on-artificial-intelligence-and-data-protection/

IDC (2019). Worldwide Spending on Artificial Intelligence Systems Will Be Nearly $98 Billion in 2023, According to New IDC Spending Guide.' International Data Corporation. Available at: https://www.idc.com/getdoc.jsp?containerId=prUS45481219

IDC (2020). IDC Forecasts Strong 12.3% Growth for AI Market in 2020 Amidst Challenging Circumstances. International Data Corporation. Available at: https://www.idc.com/getdoc.jsp?containerId=prUS46757920#:~:text=04%20Aug%202020-,IDC%20Forecasts%20Strong%2012.3%25%20Growth%20for%20AI%20Market%20in%202020,increase%20of%2012.3%25%20over%202019

IEEE (2019). Ethical aspects of autonomous and intelligent systems. Available at: http://globalpolicy.ieee.org/wp-content/uploads/2019/06/IEEE19002.pdf

IMDA, PDPC (2020). 'Model Artificial Intelligence Governance Framework Second Edition'. Available at: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

Ingold, D. and Soper, S. (2016). 'Amazon does not consider the race of its customers. Should it?' *Bloomberg*. 21 April 2016. Available at: https://www.bloomberg.com/graphics/2016-amazon-same-day/

Ipsos (2020). European enterprise survey on the Use of Technologies based on Artificial Intelligence. Report. European Commission. Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/european-enterprise-survey-use-technologies-based-artificial-intelligence

ISO (n.d.). ISO/IEC JTC 1/SC 42 Artificial intelligence. International Organisation for Standardisation. Available at: https://www.iso.org/committee/6794475.html

ITI (2017). AI Policy Principles. Information Technology Industry Council. Available at: https://www.itic.org/news-events/news-releases/iti-unveils-first-industry-wide-artificial-intelligence-policy-principles

Jagtap, R. (2020). 'A comprehensive guide to Generative Adversarial Networks (GANs)'. *Towards Data Science*. Available at: https://towardsdatascience.com/a-comprehensive-guide-to-generative-adversarial-networks-gans-fcfe65d1cfe4

Japanese Society for Artificial Intelligence (2017). Ethical guidelines. Available at: http://ai-elsi.org/archives/514

Japanese Ministry of Foreign Affairs (2019). 'G20 Ministerial Statement on Trade and Digital Economy'. Available at: https://www.mofa.go.jp/files/000486596.pdf

Jere, S., Fan, Q., Shang, B., Li, L. and Liu, L. (2020). 'Federated learning in mobile edge computing: An edge-learning perspective for beyond 5G'. *ArXiv.* Available at: https://arxiv.org/abs/2007.08030

Jernigan, Carter and Mistree, Behram F.T. (2009). Gaydar: Facebook friendships expose secual orientation, *First Monday* 14(10)

Jobin, A., Ienca, M. and Vayena, E. 'The global landscape of AI ethics guidelines'. *Nat Mach Intell* 1**,** 389–399. Available at: https://doi.org/10.1038/s42256-019-0088-2

Joyce, J. (2017). 'Trump, Twitter and his "filter bubble"'. *BBC News.* 30 November 2017. Available at: http://www.bbc.com/news/world-us-canada-42187596

JRC (2020). 'AI Watch – Artificial Intelligence in public services'. Available at: https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/ai-watch-artificial-intelligence-public-services

Kaiser, J. and Rauchfleisch, A. (2020). 'Birds of a Feather Get Recommended Together: Algorithmic Homophily in YouTube's Channel Recommendations in the United States and Germany'. Available at: https://journals.sagepub.com/doi/full/10.1177/2056305120969914

Kaissis G.A. et al. (2020). 'Secure, privacy-preserving and federated machine learning in medical imaging'. Available at: https://www.nature.com/articles/s42256-020-0186-1

Kaldestatt, O. and Myrstad, F. (2018). 'New analysis shows how Facebook and Google push users into sharing personal data.' *Forbrukerradet.* Available at: https://www.forbrukerradet.no/side/facebook-and-google-manipulate-users-into-sharing-personal-data/

Kalff, D. and Renda A. (2019). 'Hidden Treasures'. Available at: https://www.ceps.eu/ceps-publications/hidden-treasures/

Kamensky S. (2020). 'Artificial Intelligence and Technology in Health Care: Overview and Possible Legal Implications'. Available at: https://via.library.depaul.edu/cgi/viewcontent.cgi?article=1382&context=jhcl

Kamiran F. and Calders T. (2009). 'Classifying without discriminating'. Available at: https://ieeexplore.ieee.org/abstract/document/4909197/authors#authors

Kay, M., Matuszek, C. and Munson, S.A. (2015). 'Unequal representation and gender stereotypes in image search results for occupations'. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems ACM, 3819.

Kayser-Bril, N. (2020). 'Unchecked use of computer vision by police carries high risks of discrimination.' *AlgorithmWatch.* Available at: https://algorithmwatch.org/en/story/computer-vision-police-discrimination/?etcc_med=newsletter&etcc_cmp=nl_algoethik_18082&etcc_plc=aufmacher&etcc_grp

Kervizic, J. (2019). 'Overview of the different approaches to putting Machine Learning (ML) models in production'. Available at: https://medium.com/analytics-and-data/overview-of-the-different-approaches-to-putting-machinelearning-ml-models-in-production-c699b34abf86

Kim, S. (2018). 'Crashed Software: Assessing Product Liability for Software Defects in Automated Vehicles'. Available at: https://scholarship.law.duke.edu/dltr/vol16/iss1/9/

Kitchin, R. (2020). 'Civil liberties or public health, or civil liberties and public health? Using surveillance technologies to tackle the spread of COVID-19'. Available at: https://www.tandfonline.com/doi/full/10.1080/13562576.2020.1770587

Kleinberg, J., Ludwig, J., Mullainathan, S. and Sunstein, C.R. (2020). Algorithms as discrimination detectors. Available at: https://www.pnas.org/content/pnas/early/2020/07/27/1912790117.full.pdf

Koops, B.J. (2014). 'The Trouble with European Data Protection Law' Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2505692

Kouziokas, G. (2017). 'The application of artificial intelligence in public administration for forecasting high crime risk transportation areas in urban environment'. Available at: https://scholar.google.com/citations?user=YvQYVuIAAAAJ&hl=en

Kroll, J.A. et al. (2016). 'Accountable algorithms'. *University of Pennsylvania Law Review*, 165, 633-705.

Kubat, M. (2017). An introduction to machine learning. Springer International Publishing. https://doi.org/10.1007/978-3-319-63913-0_14

Latonero, M. (2019). 'Stop surveillance humanitarianism'. *New York Times.* 11 July 2019.

Lee, T. (2018). 'Report: software bug led to death in Uber's self-driving crash.' *Arstechnica.* Available at: https://arstechnica.com/tech-policy/2018/05/report-software-bug-led-to-death-in-ubers-self-driving-crash/

Lehuedé, S., Filimonov, K. and Higgins K. (2020). 'Dissent and democracy in Covid-19.' *Progressive International.* Available at: https://progressive.international/blueprint/1e766450-58f3-4ff1-8487-6ef08ee98327-lehued-filimonov-higgins-dissent-democracy-in-covid-19/en

Leufer, D. and Jansen, F. (2020). 'The EU is funding dystopian Artificial Intelligence projects.' *EURACTIV.* Available at: https://www.euractiv.com/section/digital/opinion/the-eu-is-funding-dystopian-artificial-intelligence-projects/

Levin, S. (2016). 'A beauty contest was judged by AI and the robots didn't like dark skin.' *The Guardian,* 8 September 2016. Available at: https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

Lewis, P. and McCormick, E. (2018). 'How an ex-YouTube insider investigated its secret algorithm'. Available at: https://www.theguardian.com/technology/2018/feb/02/youtube-algorithm-election-clinton-trump-guillaume-chaslot

Lievrouw, L. and Livingstone S. (2006). 'The Handbook of New Media'. Available at: https://books.google.co.uk/books?hl=en&lr=&id=9wrpPKHc0skC&oi=fnd&pg=PA1&dq=Lievrouw+and+Livingstone+(2006)&ots=G1XKN_GBAM&sig=QusIpKzas5xmPJsJ_Pd3-uDgrco#v=onepage&q=Lievrouw%20and%20Livingstone%20(2006)&f=false

LAIP (2020) *Risk Assessment and Management*, *Linking Artificial Intelligence Principles (LAIP)*. Available at: https://www.linking-ai-principles.org/term/730 (Accessed: 12 March 2021).Liu, X., Faes, L., Kale, A. et al. (2019). 'A comparison of deep learning performance against health care professionals in detecting diseases from medical imaging: a

systematic review and meta-analysis'. *Lancet Digital Health.* Available at: https://doi.org/10.1016/S2589-7500(19)30123-2

Llansó E. et al. (2020). 'Artificial Intelligence, Content Moderation, and Freedom of Expression'. Available at: https://lirias.kuleuven.be/retrieve/594053

Lombard, M. and Ditton, T. (1997). 'At the Heart of It All: The Concept of Presence'. Available at: https://academic.oup.com/jcmc/article/3/2/JCMC321/4080403

Lorenz, P. and Saslow, K. (2019a). 'Demystifying AI and AI companies.' *Stiftung Neue Verantwortung*. Available at: https://www.stiftung-nv.de/sites/default/files/demystifying_ai_and_ai_companies.pdf

Lorenz, P. and Saslow, K. (2019b). 'Artificial intelligence needs human rights.' *Stiftung Neue Verantwortung*. Available at: https://www.stiftung-nv.de/sites/default/files/ai_needs_human_rights.pdf

Lum, K. and Isaac, W. (2016). 'To predict and serve?' *Significance,* 13(5), 14–19.

Luong, B.T., Ruggieri, S. and Turini, F. (2011). 'k-NN as an implementation of situation testing for discrimination discovery and prevention'. Available at: https://dl.acm.org/doi/abs/10.1145/2020408.2020488

Macaulay, T. (2020). 'Amsterdam and Helsinki become first cities to launch AI registers explaining how they use algorithms.' *Neural.* Available at: https://thenextweb.com/neural/2020/09/28/amsterdam-and-helsinki-become-first-cities-to-launch-ai-registers-explaining-how-they-use-algorithms/

Macrae, C. (2019). 'Governing the safety of artificial intelligence in healthcare'. Available at: https://qualitysafety.bmj.com/content/28/6/495.abstract

Madhav, K.C., Shershand, S.P. and Sherchan, S. (2017). 'Association between screen time and depression among US adults.' *Prev Med Rep* 8, pp. 67-71. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5574844/

Makala, B. and Bakovic, T. (2020). 'Artificial Intelligence in the Power Sector'. Available at: https://openknowledge.worldbank.org/handle/10986/34303

Maruti Techlabs (2017). Top 10 sectors making use of Big Data analytics. *Towards Data Science.* Available at: https://towardsdatascience.com/top-10-sectors-making-use-of-big-data-analytics-be79d2301e79

Marzin, C. (2018). 'Plug and pray? A disability perspective on artificial intelligence, automated decision-making and emerging technologies.' European Disability Forum (EF). Available at: https://www.edf-feph.org/content/uploads/2020/12/edf-emerging-tech-report-accessible.pdf

Mazzucato, M. (2018). 'Mission-Oriented Research & Innovation in the European Union'. Available at: https://ec.europa.eu/info/sites/info/files/mazzucato_report_2018.pdf

McKinsey (2019). 'Confronting the risks of aritificial intelligence'. Available at: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence

McStay, A. (2020). 'Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy.' *Big Data & Society*, *7*(1). Available at: https://doi.org/10.1177/2053951720904386

METI (2019). Contract Guidelines on Utilisation of AI and Data. Ministry of Economy, Trade and Industry, Japan. Available at: https://www.meti.go.jp/press/2019/04/20190404001/20190404001-2.pdf

Misuraca, G. (2020). *AI Watch Artificial Intelligence in public services*. Science for Policy Report of the Joint Research Centre (JRC). European Commission. Available at: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120399/jrc120399_misuraca-ai-watch_public-services_30062020_def.pdf

Misuraca, G., and van Noordt, C. (2020). 'Exploratory Insights on Artificial Intelligence for Government in Europe'. Available at: https://journals.sagepub.com/doi/abs/10.1177/0894439320980449

MIT Tech Review (2019). 'The AI hiring industry is under scrutiny - but it'll be hard to fix.' *MIT Technology Review*. Available at: https://www.technologyreview.com/2019/11/07/75194/hirevue-ai-automated-hiring-discrimination-ftc-epic-bias/

Mittelstadt, B., Russell, C. and Wachter, S. (2019). 'Explaining Explanations in AI'. Available at: https://arxiv.org/abs/1811.01439

Molnar, P. and Gill, L. (2018). 'Bots at the Gate: A Human Rights analysis of automated decision-making in Canada's immigration and refugee system'. Available at: https://it3.utoronto.ca/wp-content/uploads/2018/10/20180926-IHRP-Automated-Systems-Report-Web.pdf

Monroy, M. (2019). German federal states test police software with Palantir function. Available at: https://digit.site36.net/2019/09/16/german-federal-states-test-police-software-with-palantir-function/

National Bureau of Economic Research (2001). Technology and productivity growth. *Digest,* Issue 10. Available at: https://www.nber.org/digest/oct01/technology-and-productiity-growth

Naudé, W. (2020). 'Artificial Intelligence against COVID-19: An Early Review'. Available at: https://www.iza.org/publications/dp/13110/artificial-intelligence-against-covid-19-an-early-review

Niestadt, M. (2019). 'Artificial intelligence in transport: Current and future developments, opportunities and challenges'. Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/635609/EPRS_BRI(2019)635609_EN.pdf

Nighania, K. (2018). 'Various was to evaluate a machine learning model's performance'. Available at: https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

Ntoutsi, E. et al. (2020). 'Bias in data-driven artificial intelligence systems - an introductory survey'. *WIREs Data Mining and Knowledge Discovery*. 10(3) widm.1356.

Office of the Privacy Commissioner of Canada (2017). Appearance before the Standing Committee on Access to Information, Privacy and Ethics (ETHI) on the Study of the Personal Information Protection and Electronic Documents Act (PIPEDA). Available at: https://www.priv.gc.ca/en/opc-actions-and-decisions/advice-to-parliament/2017/parl_20170216/ (archived at https://perma.cc/6VQX-Y6LW).

O'Neil, C. (2016). 'How algoriths rule our working lives'. Available at: https://www.theguardian.com/science/2016/sep/01/how-algorithms-rule-our-working-lives

Ontier (2020). 'Can artificial intelligence replace a judge in Court?'. Available at: https://uk.ontier.net/news/1584/can-artificial-intelligence-replace-a-judge-in-court/en/

OECD (2019). 'Recommendation of the Council on Artificial Intelligence'. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OECD (2020). 'Model AI Governance Framework'. Available at: https://www.oecd.ai/dashboards/policy-initiatives/2019-data-policyInitiatives-24428

Ombudsman Australia (n.d.). 'Automated decision-making better practice guide'. Available at: https://www.ombudsman.gov.au/publications/better-practice-guides/automated-decision-guide

Ortega, M. et al. (2018). 'Building safe artificial intelligence: specification, robustness, and assurance'. Available at: https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1

Ortiz Hernández, S., Garrós Font, I. and Nuria, Mª. (YYYY). 'Hacia la implantación de la inteligencia artificial en nuestro sistema judicial.' *Romera Santiago*. Revista Aranzadi Doctrinal, num.3/2020.

Owaida, A. (2020). 'Facial recognition technology banned in another US city.' *We Live Security.* Available at: https://www.welivesecurity.com/2020/06/25/boston-facial-recognition-technology-banned-another-us-city/

Özgen, A.C. and Ekenel, H. (2020). Words as art materials: generating paintings with Sequential GANs. Available at: https://www.researchgate.net/publication/342829932

Pagallo, U. and Quattrocolo, S. (2018). In W. Barfield and U. Pagallo (Eds.), *Research Handbook on the Law of Artificial Intelligence.* Edwar Elgar Publishing Limited.

Park, M. (2020). 'South Korean mother given tearful VR reunion with deceased daughter.' *Reuters,* 14 February 2020. Available at: https://www.reuters.com/article/us-southkorea-virtualreality-reunion/south-korean-mother-given-tearful-vr-reunion-with-deceased-daughter-idUSKBN2081D6

Partnership on AI (n.d.). Tenets of the Partnership on AI to Benefit People and Society. Available at: https://www.partnershiponai.org/tenets/

Pasquale, F. (2016). 'Two Narratives of Platform Capitalism'. Available at: https://ylpr.yale.edu/two-narratives-platform-capitalism

Patel, A.R. et al. (2017). 'Vitality of Robotics in Healthcare Industry: An Internet of Things (IoT) Perspective'. Available at: https://link.springer.com/chapter/10.1007/978-3-319-49736-5_5

PDPC (2020). 'Singapore's Approach to AI Governance'. Available at: https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework

Prates, M.O.R., Avelar, P.H. and Lamb, L.C. (2020). 'Assessing gender bias in machine translation: a case study with Google Translate'. *Neural Comput & Applic,* 32, 6363–6381.

Pugh, A. (2020). 'Lithuanian contact-tracing app suspended.' *Global Data Review,* 28 May 2020. Available at: https://globaldatareview.com/coronavirus/lithuanian-contact-tracing-app-suspended

Rachum-Twaig, O. (2020). 'Whose robot is it anyway?'. Availabe at: https://heinonline.org/HOL/LandingPage?handle=hein.journals/unilllr2020&div=34&id=&page=

Rahman, S., Tully, P. and Foster, L. (2019). Attention is all they need: Combating social media information operations with neural language models. Available at: https://www.fireeye.com/blog/threat-research/2019/11/combatting-social-media-information-operations-neural-language-models.html

Reardon, S. (2019). 'Rise of Robot Radiologists'. Available at: https://go.gale.com/ps/anonymous?id=GALE%7CA649636850&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=00280836&p=AONE&sw=w

Redman, T. (2016). 'Bad Data Costs the U.S. $3 Trillion Per Year'. Available at: https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

Reding (2012) in Mordini and Tzovaras. 'Second Generation Biometrics: The Ethical, Legal and Social Context'. ISBN 978-94-007-3892-8

Reeves, B. and Nass, C. (1996). 'The Media Equation How People Teat Computers, Television, and New Media Like Real People and Platforms'. Available at: https://www.researchgate.net/publication/37705092_The_Media_Equation_How_People_Treat_Computers_Television_and_New_Media_Like_Real_People_and_Pla

Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). Algorithmic impact assessments: a practical framework for public agency accountability. AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf

Renda, A. (2015). 'Searching for harm or harming search? A look at the European Commission's antitrust investigation against Google'. Available at: http://aei.pitt.edu/67571/1/AR_Antitrust_Investigation_Google.pdf

Renda, A. (2018). 'The legal framework to address "fake news": possible policy actions the EU level' In-Depth Analysis Requested by the IMCO Committee. Available at: http://aei.pitt.edu/94231/1/AR_FakeNews_IMCO.pdf

Renda, A. (2019). 'Artificial Intelligence Ethics, governance and policy challenges'. Available at: https://www.ceps.eu/wp-content/uploads/2019/02/AI_TFR.pdf

Renda, A., Schrefler, L., Luchetta, G. and Zavatta, R. (2013). *Assessing the costs and benefits of regulation*. Study for the European Commission, Final Report for the European Commission, Secretariat-General.

Rich, E. and Knight, K. (1991). 'Artificial Intelligence'. Available at: https://books.google.co.uk/books/about/Artificial_Intelligence.html?id=6P6jPwAACAAJ&redir_esc=y

Richardson, R., Schultz, R. and Crawford, K. (2019). 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice'. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423

Rocher, L., Hendrickx, J. and Montjoye, Y-A. (2019). 'Estimating the success of re-identifications in incomplete datasets using generative models'. *Nature Communications,* 10. 10.1038/s41467-019-10933-3.

Ronsin, X. and Lampos, V. (2018). 'Appendix I – In-depth study on the use of AI in judicial systems, notably AI applications processing judicial decisions and data'. In: European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their

environment. Strasbourg, CEPEJ - Commission Européene pour l'Efficacité de la Justice. Available at: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c

Ross, C. and Swetlitz, I. (2018). 'IBM's Watson supercomputer recommended "unsafe and incorrect" cancer treatments, internal documents show'. *Statnews.* Available at: https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

Roth, A. and Dwork, C. (2013). 'The algorithmic foundations of differential privacy.' *Found. Trends Theoretical Comp. Sci.* 9, 211–407.

Russell, S. and Norvig, P. (2009). *Artificial Intelligence: A modern approach.* (3rd ed.). Pearson. Sareen, Saltelli and Rommetveit (2020). 'Ethics of quantification: illumination, obfuscation and performative legitimation'. Available at: https://www.researchgate.net/publication/338959066_Ethics_of_quantification_illumination_obfuscation_and_performative_legitimation

Sánchez-Monedero, J., Dencik, L. and Edwards, L. (2020). 'What does it mean to "solve" the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems'. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 458-468.

Scharkow, M., Mangold, F., Stier, S. and Breuer, J. (2020). 'How social network sites and other online intermediaries increase exposure to news.' Proceedings of the National Academy of Sciences of the United States of America, 11 February 2020. Available at: https://urldefense.com/v3/__http:/x3ysn.mjt.lu/lnk/AL8AAGV0558AAAACu1YAAADMZWkAAAAAKxoAAB1gABB0KQBeT483m39wgiP1QfSZwJ5hMyWyfwAQJ0I/14/KcZ72PnKRoI-OxbNgB7X6w/aHR0cHM6Ly93d3cucG5hcy5vcmcvY29udGVudC8xMTcvNi8yNzYx__;!!DOxrgLBm!S4LfHhBbZfeQIUaQLuyfzqbsyWr6ROn5ZI2CxS6MkgB6iljylFjGYbcQVXEGcAHHLvDE$

Schmelzer, R. (2020). 'Machine learning limitations marked by data demands'. Available at: https://searchenterpriseai.techtarget.com/feature/Machine-learning-limitations-marked-by-data-demands

Schulz, W. et al. (2017). Algorithms and Human Rights. Study on the Human Rights Dimensions of Automated Data Processing Techniques (in particular algorithms) and Possible Regulatory Implications, Council of Europe Study DGI(2017)12

Scott, P.J. and Yampolskiy, R.V. (2019). 'Classification Schemas for Artificial Intelligence Failures'. Available at: https://delphi.lexxion.eu/article/DELPHI/2019/4/8

Serrato, R. (2018). 'How YouTube's algorithm amplified the right during Chemnitz: algorithmic accountability.' *AlgorithmWatch*. Berlin.

SIA Partners (2017). Insights: Preparing for the GDPR. SIA Partners. Available at: https://sia-partners.co.uk/preparing-gdpr-need-15m-300-450-per-employee-average-implement-gdpr/

Sharma, G.D. et al. (2020). 'Artificial intelligence and effective governance: A review, critique and research agenda' Available at: https://www.sciencedirect.com/science/article/pii/S2666188819300048

Sheshasaayee, A. and Bhargavi, K. (2017). 'A study of automated decision-making systems'. *International Journal of Engineering and Science,* 7(1), 28-31. http://www.researchinventy.com/papers/v7i1/E07012831.pdf

Simonite, T. (2020). 'The world has a plan to rein in AI - but the US doesn't like it.' *Wired.* 1 June 2020. Available at: https://www.wired.com/story/world-plan-rein-ai-us-doesnt-like/

Smith, B.W. (2020). 'How Reporters Can Evaluate Automated Driving Announcements' Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3747036

Song, C. et al. (2017). 'Machine Learning Models that Remember Too Much'. Available at: https://dl.acm.org/doi/abs/10.1145/3133956.3134077

Spiekermann, M. (2019). 'Data marketplaces: trends and monetisation of data goods'. *Intereconomics*, 54, 208-216.

Standards Australia (2020). 'An Artificial Intelligence Standards Roadmap: Making Australia's Voice Heard'. Available at: https://www.standards.org.au/getmedia/ede81912-55a2-4d8e-849f-9844993c3b9d/1515-An-Artificial-Intelligence-Standards-Roadmap12-02-2020.pdf.aspx

Standing Committee on Access to Information, Privacy and Ethics (2018) *Towards Privacy by design: Review of the Personal Information Protection and Electronic Documents Act*. Standing Committee on Access to Information, Privacy and Ethics.Stanley, J. (2018). The problem with using face recognition on fans at a Taylor Swift concert. ACLU Speech, Privacy and Technology Project. Available at: https://www.aclu.org/blog/privacy-technology/surveillance-technologies/problem-using-face-recognition-fans-taylor-swift

Statista (2019). Revenues from the artificial intelligence software market worldwide from 2018 to 2025, by region. Available at: https://www.statista.com/statistics/721747/worldwide-artificial-intelligence-market-by-region/

Statt, N. (2019). Google and DeepMind are using AI to predict the energy output of wind farms. Available at: https://www.theverge.com/2019/2/26/18241632/google-deepmind-wind-farm-ai-machine-learning-green-energy-efficiency

Stiftung Neue Verantwortung (2020). Mapping AI Governance Fora.

Strubell, E., Ganesh, A. and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. July 2019. *ARXIV.* Available at: https://arxiv.org/abs/1906.02243

Sullivan, H.R and Schweikart, S.J. (2019). 'Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI?'. Available at: https://journalofethics.ama-assn.org/article/are-current-tort-liability-doctrines-adequate-addressing-injury-caused-ai/2019-02

Sunstein, C.R. (2001). *Republic.com*, Princeton University Press.

Supreme Court of Wisconsin, State of Wisconsin v. Eric L. Looms, pag. 49 par. 99. 2016.

Sweeney, L. (2013). Discrimination in Online Ad Delivery, *Communications of the ACM* 56(5), p. 44-54.

The Economist (2019). 'How the world will change as computers spread into everyday objects'. Available at: https://www.economist.com/leaders/2019/09/12/how-the-world-will-change-as-computers-spread-into-everyday-objects

The Guardian (2018). 'NTSB 'unhappy' with Tesla for releasing information about fatal crash'. Available at: https://www.theguardian.com/technology/2018/apr/02/ntsb-unhappy-tesla-fatal-crash-autopilot

The Law Society of England and Wales. (2019). Algorithms in the Criminal Justice System. Available at: https://www.ailira.com/wp-content/uploads/2019/06/algorithms-in-criminal-justice-system-report-2019.pdf

The Public Voice (2018). Universal guidelines for artificial intelligence. Available at: https://thepublicvoice.org/ai-universal-guidelines/

Themistocleous, M., Papadaki, M., and Kamal, M. M. (Eds.). (2020). Information Systems: 17th European, Mediterranean, and Middle Eastern Conference, EMCIS 2020, Dubai, United Arab Emirates, November 25–26, 2020, Proceedings. Springer International Publishing. https://doi.org/10.1007/978-3-030-63396-7

Thiel, V. (2019). 'Defective computing: How algorithms use speech analysis to profile job candidates.' *AlgorithmWatch.* Available at: https://algorithmwatch.org/en/story/speech-analysis-hr/

Treudeau, J. (2018) 'Mandate for the International Panel on Artificial Intelligence'. Available at: https://pm.gc.ca/en/news/backgrounders/2018/12/06/mandate-international-panel-artificial-intelligence

Tully, P. and Foster, L. (2020). 'Repurposing Neural Networks to Generate Synthetic Media for Information Operations'. *FireEye.* Available at: https://www.fireeye.com/blog/threat-research/2020/08/repurposing-neural-networks-to-generate-synthetic-media-for-information-operations.html

Turner Lee, N., Resnick, P. and Barton, G. (2019). 'Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms'. *Brookings*. Available at: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Tuvsud (n.d.). MDR conformity assessment procedures. Available at: https://www.tuvsud.com/en/industries/healthcare-and-medical-devices/medical-devices-and-ivd/medical-device-market-approval-and-certification/medical-device-regulation/mdr-conformity-assessment-procedures

UK Office for Artificial Intelligence (2019). Guidance: Assessing if artificial intelligence is the right solution. UK Government. Available at: https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution#consider-your-current-data-state

UK Government (2019). Guidance on building and using artificial intelligence in the public sector. Available online: https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector

UNESCO (2017). Report of COMEST on Robotics Ethics, World Commission on the Ethics of Scientific Knowledge and Technology. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000253952

UNI Global Union (2017). 10 principles for ethical AI. Available at: http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/

United Nations AI for Good (2019). Global summit. 28-31 May 2019. Geneva, Switzerland. Available at: https://aiforgood.itu.int/2019-event/

UN General Assembly (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UN Doc. A/73/348, para 19.

Urban, Karaganis and Schofield (2016). 'Notice and Takedown in Everyday Practice'. Available at: https://www.copyrightevidence.org/wiki/index.php/Urban,_Karaganis_and_Schofield_(2016)

US Association for Computing Machinery (2017). Statement on algorithmic transparency and accountability. Available at: https://www.acm.org/articles/bulletins/2017/january/usacm-statement-algorithmic-accountability

- Van Demark, D. (2020) *The Future of AI Regulation: The Government as Regulator and Research & Development Participant*. Available at: https://mcdermott-will-emery-2793.docs.contently.com/v/the-future-of-ai-regulation-the-government-as-regulator-and-research-development-participant1 (Accessed: 12 March 2021).

Vaas, L. (2019). 'Emotion detection should be regulated, AI Now says.' *Naked Security by SOPHOS.* Available at: https://nakedsecurity.sophos.com/2019/12/16/emotion-detection-in-ai-should-be-regulated-ai-now-says/

Van der Marel, E., Bauer, M., Lee-Makiyama, H. and Verschelde, B. (2016). 'A methodology to estimate the costs of data regulations'. *International Economics*, *146*, 12-39.

Varghese, S. (2019). 'The junk science of emotion-recognition technology.' *The Outline*. Available at: https://theoutline.com/post/8118/junk-emotion-recognition-technology?zd=1&zi=xmlnbkbj

Villasenor, J. (2019). 'Artificial intelligence and bias: Four key challenges'. Available at: https://www.brookings.edu/blog/techtank/2019/01/03/artificial-intelligence-and-bias-four-key-challenges/

Vought, R.T. (2019). 'Memorandum for the Heads of Executive Departments and Agencies'. Available online: https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

Vucheva, M., Rocha, M., Renard, R. and Stasinopolous, D. (2020). *Study on the use of innovative technologies in the justice field*. European Commission. Brussels. Available at: https://op.europa.eu/en/publication-detail/-/publication/4fb8e194-f634-11ea-991b-01aa75ed71a1/language-en

Wachter, S. (2019). 'Data Protection in the Age of Big Data'. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3355444

Wachter, Mittelstadt and Russel (2020). 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI'. Available at: https://arxiv.org/abs/2005.05906

Weiss, M. (2019). 'Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions.' *TechScience.* Available at: https://techscience.org/a/2019121801/

West, D. (2018). 'Brookings survey finds worries over AI impact on jobs and personal privacy, concern U.S. will fall behind China.' *Brookings,* 21 May 2018. Available at: https://www.brookings.edu/blog/techtank/2018/05/21/brookings-survey-finds-worries-over-ai-impact-on-jobs-and-personal-privacy-concern-u-s-will-fall-behind-china/

Williams, P. (2018). 'BEING MATRIXED: THE (OVER)POLICING OF GANG SUSPECTS IN LONDON'. Available at: https://www.stop-watch.org/uploads/documents/Being_Matrixed.pdf

Wirtz, B.W. et al. (2019). 'Artificial Intelligence and the Public Sector—Applications and Challenges'. Available at: https://www.tandfonline.com/doi/abs/10.1080/01900692.2018.1498103

Wybitul, T. and Brams, I. (2020). 'LG Darmstadt: 1.000 Euro immaterieller Schadensersatz für Datenschutzverstoß'. *Latham Germany.* Available at: https://www.lathamgermany.de/2020/11/lg-darmstadt-1-000-euro-immaterieller-schadensersatz-fur-datenschutzverstos/

Yampolskiy, R. (2016). 'Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures'. Available at: https://www.researchgate.net/publication/309424933_Artificial_Intelligence_Safety_and_Cybersecurity_a_Timeline_of_AI_Failures

Zhang, M. (2015). 'Google photos tags two African Americans as gorillas through facial recognition software.' *Forbes.* Available at: https://www.forbes.com/sites/mzhang/2015/07/01/google-photostags-two-african-americans-as-gorillas-through-facial-recognitionsoftware/#55d05821713d

Zheng, M. et al. (2020). 'How causal information affects decisions'. Available at: https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-020-0206-z

- Zhu, J. (2018) 'Canada Treasury Board's Directive on Automated Decision-Making', *Laboratoire de cyberjustice*. Available at: https://www.cyberjustice.ca/2018/11/25/canada-treasury-boards-directive-on-automated-decision-making/ (Accessed: 9 March 2021).

Zhu, X. (2005). Semi-supervised learning literature survey (Technical Report). University of Wisconsin-Madison Department of Computer Sciences. Available at: https://minds.wisconsin.edu/handle/1793/60444

Zisov, N. and Targov, T. (n.d.). 'Risks of Artificial Intelligence – Safety Aspects'. Available at: https://boyanov.com/risks-of-artificial-intelligence-safety-aspects/

Zilobate, I. and Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models, *Artificial Intelligence and Law* 24, p. 183-201.

Zuboff, S. (2018). 'Surveillance Capitalism and the Challenge of Collective Action'. Available at: https://journals.sagepub.com/doi/full/10.1177/1095796018819461

Zuiderveen Borgesius, F.J. (2018). 'Discrimination, artificial intelligence and algorithmic decision-making'. Available at: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73

Zuiderveen Borgesius, F.J. (2020). Strengthening legal protection against discrimination by algorithms and artificial intelligence, The International Journal of Human Rights 24(10), p.1572-1593.

Zuiderveen Borgesius, F.J. et al. (2016).'Should we worry about filter bubbles?'. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2758126

ANNEX 1: SUMMARY OF AI RISKS TO FUNDAMENTAL RIGHTS

Table 43 - *Evidence cases of AI systems posing (long-term) risks to fundamental rights*[89]

Case	Long-term risk	Affected entity	Origin and use of data	Degree of intervention
I DIGNITY (EU Charter of Fundamental Rights Articles 1-5: **Human dignity, Right to life, Right to the integrity of the person, Prohibition of torture and inhuman or degrading treatment or punishment, Prohibition of slavery and forced labour**)				
Personalisation ● The 'Blue Feed, Red Feed' project shows how Facebook feeds depict different realities based on their political predispositions. ● Algorithm-based user feedback is used to build highly personalised feeds that create immersive media environments, causing addiction and opinion manipulation for users. The research claims that the 'algorithm itself should be considered an immersive media environment that permits users to consume unique media feeds that may affect civic action's. ● In the digital 'attention economy', technologies compete to capture and exploit attention, rather than supporting individual true goals. Intentional persuasive design goals of social media, and digital technologies in general, instead leads to addictive behaviour. ● The Tinder scoring algorithm compares users and matches people who have similar levels of 'desirability': Users with less successful matching requests will likely never get to swipe on profiles clustered in the more successful ranks. ● TikTok uses facial recognition to analyse profile pictures for recommending new accounts based on the physical appearance of the people a user already follows. ● Personalisation and targeted content in the form of 'dark ads' are part of the information systems that people use to process news, e.g. on Facebook. ● A recent report on AI in the advertising industry notes consumer harms because AI enables the excessive collection of data, restricts choices - leading to discrimination - contributes to the manipulation of and harm to vulnerable people, and fuels online scams.	High long-term risks to opinion plurality and the right to mental safety and integrity.	B2C	Data are mostly voluntarily provided because users opt-in to the service/networks. However, no alternative networks in place. Mostly, no possibility to opt-out as data is captured by the use.	Low/no degree of intervention for users. No traceability of data (repurposing).
Erosion of human agency – Several studies confirm that it is impossible for researchers to fully protect real identities in datasets. – Insufficient means, practical tools or applications are available to users to provide meaningful consent. – Online advertising industry leaves users with little control over their data. Although collective redress can be sought under the GDPR, the complexity of the system means that	High long-term risks to eroding privacy and human agency in the digital media environment.	B2C	Voluntarily provided by users.	Little to medium degree of intervention for regular digital media use, dependent on service.

[89] The table above presents a non-exhaustive overview of documented cases in which AI has been (partly) responsible for fundamental rights violation. The cases are grouped into the categories I – VI (in blue), which correspond to the EU Charter. The majority of cases do not exclusively belong to one category as they may pose risks to multiple EU fundamental rights. The table above also includes examples of critical AI systems operating outside of the EU legislative scope as these are increasingly pervasive and transcend national borders.

consumers may not even understand they have been discriminated against or had their rights impinged. New formats, technologies and opportunities to engage are increasing the likelihood of bombardment and the prevalence of unreliable or biased AI gives cause for concern. – AI shapes immersive media environments, fostering short-term engagement.				
Profiling ● Big Data analytics and AI draw non-intuitive and unverifiable inferences and predictions about the behaviours, preferences and private lives of individuals, who are granted little control or oversight over how their personal data are used to draw those inferences. ● Online platform providers use behavioural advertisement and can infer very sensitive information (e.g. ethnicity, gender, sexual orientation, religious beliefs) about individuals to target or exclude certain groups from products and services, or to offer different prices.	High long-term risks to the right to non-discrimination and privacy.	Mainly B2C	Data are involuntarily provided/used to assess specific consumer patterns.	No degree of intervention as customers are often unaware that they are being profiled/offered discriminatory pricing.
Nudging ● An Instagram analysis presents strong evidence that pictures showing more skin are shown to users more often than pictures that do not. Sexually suggestive images, as well as nudity from either gender, appeared significantly more often on data donors' newsfeeds than in the posts created by monitored accounts. ● A report on online manipulation and online harm analysed the use of nudges in digital markets. It found that consumer biases such as cognitive limitations or psychological weaknesses are often exploited. ● Netflix is alleged to experiment with the order in which episodes are listed, based on the inferred sexuality of users and the corresponding storyline (homosexual or heterosexual characters). ● Google Shopping showed its own comparison-shopping website on Google's search engine platform in a more prominent way than similar services by providing a design that will exploit inertia to nudge users to use another service provided by Google. ● AI systems are increasingly central in shaping and manipulating consumer behaviour. ● The US Military is studying and using data-driven social media propaganda to manipulate news feeds to change perceptions of military actions.	High long-term impact on psychology and behavioural traits of individuals on social media.	B2C	Data are involuntarily provided by using networks/services.	Little to no degree of intervention as nudging often happens unconsciously.
Emotion recognition ● A US university considered using a system based on Microsoft's facial recognition and affect detection tools to observe students in the classroom using a webcam. The system predicts the students' emotional state. An overview of student sentiment is viewable by the teacher, who can then shift their teaching in a way that 'ensures student engagement', as judged by the system. ● In the UK, facial recognition technology enabling people's moods to be picked up by CCTV is set to be trialled. The software can detect people wearing hats and glasses and claims to find people showing a certain mood or expression. ● A 2018 testing of two mental health chatbots by the BBC revealed that the applications failed to properly handle children's reports of sexual abuse, even though both apps were considered suitable for children. ● https://nakedsecurity.sophos.com/2019/12/16/emotion-detection-in-ai-should-be-regulated-ai-now-says/	High long-term impact on psychology and behavioural traits (see above).	B2C	Data are mostly involuntarily provided either through use or other means (e.g. cameras).	Little to no degree of intervention as emotion recognition and data collection often happens unconsciously.

• https://theoutline.com/post/8118/junk-emotion-recognition-technology?zd=1&zi=xmlnbkbj • https://www.psychologicalscience.org/publications/emotional-expressions-reconsidered-challenges-to-inferring-emotion-from-human-facial-movements.html				
II FREEDOMS EU Charter of Fundamental Rights Articles 6-19. In particular: Right to liberty and security, Respect for private and family life, Protection of personal data, Freedom of thought, conscience and religion, Freedom of expression and information, Right to education, Right to asylum.				
Demographic data collection • Facebook uses AI to map most of the population of the African continent, combining computer vision techniques, population data, and high-resolution satellite imagery to search for built-up structures across the continent. They then created population density maps based on the number of buildings observed. Eventually, the company plans to map population density around the world. • Facebook's project Aria uses AR headsets aiming to create a live, 3D map of the world, constantly updated and refreshed by people walking around with AR headsets. All data are owned by the company. While the maps of public places are publicly viewable, users' homes and belongings are private.	High risks to the rights of freedom, data protection and the right to private life. Disproportionate powers to private entities collecting data on citizens.	B2C, less B2B	Less data from citizens involved/collected, but little to no means to opt-out if people are being registered by augmented reality (AR).	Degree of intervention by governments unclear.
Data breaches • Internet-connected CloudPets toys exposed two million voice recordings, emails and other sensitive data of children and adults.	High risk to the rights of freedom, data protection and the right to private life.	B2C, less B2B	Data are always involuntarily provided due to breaches and intrusion.	No degree of intervention as citizens are unaware of their data being stolen.
Facial recognition in public spaces • In Madrid, Spain, a facial recognition system at the South Station automatically matches faces against a database of suspects and shares that information with Spanish police. • An algorithm developed by IBM using New York Police Department surveillance footage lets police search by skin colour. • PimEyes analyses face images published on social media and other internet websites for individual characteristics and stores the biometric data. The database is said to contain over 900 million faces.	Very high long-term impact on freedom, autonomy and privacy.	G2C	Data are almost always involuntarily provided due to pre-installed technology.	Almost no degree of intervention due to instalment of technology in public spaces.
Commercial data repurposing • Data repurposing by machine learning algorithms that can leak significant amounts of data. Personal information is used for their training, leading to further availability of personally identifiable data (Song et al., 2017; Shokra et al., 2017). • Companies such as DataSift take data from Twitter, Facebook and other social media and make it available for analysis for marketing and other purposes.	High long-term risks to the right to non-discrimination and privacy.	All entities affected.	Data are involuntarily provided/used to train new algorithms based on previous data/behaviour.	No degree of intervention as citizens are often unaware that they are profiled and their data used further.

• Geotagged photos on Flickr, together with the profiles of contributors, have been used as a reliable proxy for estimating visitor numbers and origins at tourist sites.				
Privacy-intrusive technologies • Amazon employs thousands of contract workers in Boston, Costa Rica, India, Romania and other countries to annotate audio recordings each day from devices powered by its assistant. • In Poland, photos and films of speed cameras and films are fed into a central processing system and automatically merged with personal data by a company. • Real-time bidding operates behind the scenes on websites and apps. It constantly broadcasts private internet consumption behaviour and location data to numerous companies. For example, Google's RTB system sends personal datasets to 968 companies. • A unique gait analysis through video data can be associated with identity, allowing for real-time tracking. Gait analysis tracking is already used in China.	High risk to the erosion of non-discrimination, data protection and the right to private life.	B2C, less B2B	Data are mostly involuntarily provided through third party tracking entities.	Low to no possibility of intervention if data collection is automated.
Lending and credit scores • In India, one primary criterion for evaluation is social media and the various data points these platforms provide, for example, on a person's political activity. • In Kenya, the company Safaricom bases its lending algorithms as 'an ambitious effort to track everyday behavior and social relations'. • In the EU, businesses lend at higher rates to borrowers with poor credit records, also known as 'sub-prime lending'.	High, sustained impairment of the living standards of future generations.	G2C, less B2C	Data can be both voluntarily and involuntarily provided.	Low to no possibility of intervention if credit scoring is automated.
Online content moderation • AI techniques (natural language processing and image recognition) in content moderation often entail false positives and false negatives; potential bias and algorithmic discrimination; large-scale processing of user data and profiling; and presumptions of appropriateness of prior censorship decisions. • Content moderating staff suffer multiple psychological problems, such as post-traumatic stress disorder (PTSD). Content moderators for Facebook working in the EU were forced to sign a form acknowledging that the work can lead to PTSD.	High long-term risks to opinion plurality, the right to information and freedom of expression.	B2C	Data are mostly voluntarily provided because users opt-in to the service/networks. However, no alternative networks in place.	Low/no degree of intervention for users. No traceability of data (re-purposing).
AI and asylum • iBorderCTRL, a Horizon 2020-funded project, aimed to create an automated border security system to detect deception based on facial recognition technology and the measurement of micro-expressions. • The company ETS tried to identify immigration fraud using voice recognition software in the UK, resulting in cancelling thousands of visas and deporting people in error.	High long-term risks and impacts to the right to a fair trial and non-discrimination.	G2C	Data are mostly involuntarily collected by migrants by requiring them to undertake video scans.	No degree of intervention for affected people (migrants). Little degree of intervention by border protection authorities, depending on use, with HITL/HOTL providing higher degrees of intervention.

Lie detection software • In the UK, Northumbria police are carrying out a pilot scheme that uses EyeDetect to measure the rehabilitation of sex offenders. • In the US, similar systems such as *SilentTalker, EyeDetect* and *Discern* are being trialled privately or by public administrations, claiming to detect lies by measuring facial expressions.	High long-term risk and impact on the right to a fair trial and non-discrimination.	G2C	Data are voluntarily or involuntarily gathered through video scans.	Little to no means of intervention if a suspected criminal is required to undertake a scan by police.
III EQUALITY EU Charter of Fundamental Rights Articles 20-26. In particular: Equality before the law, Non-discrimination, Cultural, religious and linguistic diversity, Equality between women and men, Rights of the child, Rights of the elderly, Integration of persons with disabilities				
Discrimination against minority groups • In the EU, partially abled people experience significant disadvantages in accessibility of digital services and universal design. • AI may exacerbate healthcare inequalities, in particular structural racism between ethnoracial groups, as it is a necessary feature of personalised medicine due to the increasing availability of big health data sources. • In China, Hikvision has marketed an AI camera that automatically identifies Uyghurs. This AI technology allows the PRC to automatically track Uyghur people, one of the world's most persecuted minorities.	High (erosion of non-discrimination, privacy and equal treatment of all citizens).	B2C	Data are involuntarily provided/used to train new algorithms.	No intervention.
Discrimination against ethnically diverse groups • A Google subsidiary company analyses blobs of text and produces a measure of toxicity. Texts that contained the phrase 'as a Black person' or 'as a gay person' were much more likely to be considered toxic than equivalent sentences that used other adjectives. • Leading facial recognition software (serving police in the US, Australia, and France) matches different black women's faces less reliably (more false matches) than those of white women, or black or white men. • Facial recognition technology that is trained on and tuned to Caucasian faces systematically misidentifies and mislabels individuals of other races, with significantly higher error rates. • Twitter automatically crops pictures to focus on their most 'salient' parts. People of colour are often cropped away - but not white people. • In the US, a person of colour living in a Detroit suburb was wrongfully arrested because facial recognition software used by Michigan State Police misidentified the individual. • The majority of facial recognition algorithms tested by NIST perform worse on Black, Asian, and Native American faces, and show bias against women, the elderly, and children.	High (erosion of non-discrimination, privacy and equal treatment of all citizens).	B2C	Algorithms are trained with incorrect or low-quality data.	Little to no degree of intervention.

Recruitment process • Algorithm-based selection processes ('robo-recruiting') search applicant profiles for specific qualifications and keywords. However, the applicants are rarely informed when they are evaluated automatically. • Amazon deployed an AI recruiting tool that showed bias against women: the software concluded that men should be preferred over women when it comes to filling job vacancies. The software also reproduced other discriminating selection criteria. • Bias is introduced, replicated and hidden by automated hiring systems. • The report explores how predictive tools affect equity throughout the entire hiring process, explores popular tools that many employers currently use, concluding that without policy intervention, bias will arise in predictive hiring tools by default. • The IBM Watson Personality Insights service scrapes social media, enterprise data, or other digital communications (email, text messages, tweets, and forum posts) and combines these data with 'linguistic analytics to infer individuals' intrinsic personality characteristics'. The tool claims to 'determine individuals' consumption preferences, which indicate their likelihood to prefer various products, services, and activities'. • In Austria, a software extension started automatically evaluating the chances of a job placement on the basis of a statistical model, which may impact the granting or denial of funds.	High (potential to normalise discriminatory hiring practices).	All.	Voluntarily provided by users (job seekers), data used to automate tasks and to train new algorithms.	Little means of intervention if user wants to be hired.
Education • In the EU, a predictive algorithm assigned final grades for the International Baccalaureate without explanation or means for meaningful redress. Serious mismatches emerged between expected grades based on a student's prior performance, and those awarded by the algorithm. In some cases, the unexpectedly poor grades generated resulted in scholarships and admissions offers being revoked. • In the UK, students' exam results based on a controversial algorithm were alleged to be biased against students from poorer backgrounds. • In the UK, thousands of students were wrongly and forcibly deported based on a flawed algorithmic assessment of their English proficiency exams.	High long-term impact for autonomous learning and assessment of students.	B2C	Data assembled from previous records, no degree of intervention by students.	None.
Linguistic diversity • Google Translate almost always changes the gender of occupations to fit gross stereotypes for translations between EU langages (e.g. 'Der Krankenpfleger' (the male nurse in German) becomes 'l'infirmière' (the female nurse) in French). • Only a fraction of global languages are supported by virtual personal assistants, predictive text, and speech recognition and machine translation tools: Apple's Siri supports 21 languages, Amazon's Alexa eight, and Google Home 13. Google Translate supports 108 languages out of 7,117 known living languages worldwide. The language in which a service is available 'profoundly impacts a community's access to technology and the prevalence of a language's everyday use'.	High risk of discrimination and erosion of minority languages in the EU.	All.	Data are gathered via existing linguistic databases online.	No degree of language choice if relevant language is not offered.
IV SOLIDARITY EU Charter of Fundamental Rights Articles 27-38. In particular: Workers' right to information and consultation within the undertaking, Right of collective bargaining and action,				

Fair and just working conditions, Family and professional life, Social security and social assistance, Healthcare, Consumer protection				
Housing • Facebook engages in housing discrimination with its advertising practices. • Algorithms discriminate against women and older workers for housing and employment adverts on Facebook.	High (potential to create distrust in public welfare and services).	G2C	Often, the data are publicly available, or users provide the data in need of services.	Little degree of intervention if user is dependent on service.
Healthcare • In Denmark, an automated risk assessment experiment in the field of social welfare is a project that measures chronically ill patients' behaviour in order to estimate when or how further efforts are necessary, namely whether patients should be admitted to hospital with severe conditions. • IBM's Watson recommended unsafe and incorrect cancer treatments. • In the US, a healthcare algorithm affecting millions is biased against black patients. • China's largest insurer, Ping An, has apparently started employing facial recognition to identify untrustworthy and unprofitable customers.	High (potential to create distrust in public healthcare system and non-discrimination as well as in public authorities).	G2C	Data are already with healthcare providers/public administrations, users cannot claim any data ownership or agency. Data use is less comprehensible by users.	Low degree of intervention by users. Low-medium degree of intervention if the AI system is used to evaluate large-scale numbers of cases. Higher degree of intervention if AI system is used with HITL/HOTL for evaluating single cases.
Fraud risk assessment • In the Netherlands, the SyRI model was used to determine the risk of fraud in the area of social security, income-dependent schemes, taxes and social security, and labour laws.	High long-term risk to the right to non-discrimination and the right to privacy.	G2C, less B2C	Data are mostly involuntarily provided.	Low to no possibility of intervention if fraud risk determination is automated.
Child welfare • In the UK, the Gladsaxe case used a tracing tool as part of the country's ghetto plan in January 2018 to detect children in vulnerable circumstances at an early stage. Municipalities were allowed to collect and combine information on children from different public sources and to categorise it according to specific 'risk indicators'. • In Wrocław, Poland, an algorithm automatically qualified children for individual nurseries and placed them into appropriate groups based on data from parents' declarations. The system wrongly left out children in a certain age group. • A paper found that the use of predictive analytics in child welfare may result in problems related to cognitive biases, previous marginalisation data and structural disparities. • In France, allocation committees for places in public daycare facilities are increasingly replaced by algorithms that do not always consider individual factors and specific situations. • In the UK, none of 32 tested predictive models for life trajectories met the threshold that was set in advance for success, with most falling far short. The algorithmic models attempted to predict children's futures based on real-world data from four UK communities.	High long-term risk to the right to non-discrimination and the right to privacy.	G2C	Data are involuntarily provided.	No possibility of intervention for children to object to their data being analysed.

Social welfare • The Polish Ministry of Labour and Social Policy implemented a system based on profiling the unemployed to decide how to distribute labour market programmes. An algorithm scored people based on data from computer-based interviews and 24 personal data points. • In Spain, an app/algorithm determines whether at-risk citizens are entitled to a discount on energy bills. The app is reported to result in lengthy and complex application procedures, with dozens of applicants wrongly dismissed.	High long-term risk to the right to good administration and right to social benefits.	G2C	Data are mostly involuntarily provided.	No possibility of intervention for citizens to object to their data being analysed.
Price discrimination • Angwin et al. (2015) found that the company's price differentiation practice led to higher prices for people with an Asian background: 'Customers in areas with a high density of Asian residents were 1.8 times as likely to be offered higher prices, regardless of income'. • In Germany, contract data from as many customers as possible should be stored by Schufa and a Munich credit agency to prevent electricity and gas customers from changing providers frequently. Electricity and gas companies could use such databases to see customers who have changed frequently and could either systematically reject them or withhold attractive conditions.	High long-term risk to the right to non-discrimination and the right to privacy.	Mainly B2C, less B2B	Data are involuntarily provided/used to assess specific consumer patterns.	No degree of intervention as customers are often unaware that they are being profiled/offered discriminatory pricing.
Surveillance at work • The article gives an overview of technological advancements that enable surveillance within and outside the workplace, and the practices blend into private lives. • A former employee of a money transfer firm says she was told to keep her phone on at all times and was dismissed weeks after being 'scolded' for uninstalling the app. • Amazon automatically generates any warnings or terminations regarding quality or productivity without input from supervisors. • Applications attempt to increase 'employee performance monitoring based on various data samples that are generated in the course of everyday processes within the company. Other products offer procedures for continuous staff surveys in order to analyse team dynamics and the job satisfaction of individual employees'.	High long-term risk to the right to privacy and the right to non-discrimination.	Mainly B2C and B2B	Involuntarily provided data by employers, used to increase efficiency.	Almost no degree of intervention by employees.
Workers' rights • ADM is used to allocate employees, tasks and shifts, sometimes resulting in unfair procedures. For example, Foodora workers are allowed to choose shifts depending on their effectiveness rating. • Digital platform operators (e.g. Uber, Foodora, Helpling) use apps to replace management staff by automating order allocation and performance control. Little to no legal means are available to freelance workers against the automated decisions of the apps/systems.	High (erosion of non-discrimination, privacy and the right to workers' collective bargaining and action)	Mainly B2C and B2B	Involuntarily collected and provided data as part of the job performance.	Almost no degree of intervention by employees.
V CITIZENS' RIGHTS EU Charter of Fundamental Rights Articles 39-46. In particular: Right to good administration, Freedom of movement and of residence				

Scoring in public administration • In Trelleborg, Sweden, an algorithm collects data from several databases (tax agency, bureau for housing support, etc.) and decides whether or not applicants can receive social benefits. • In Denmark, a point system was designed (but not deployed) to detect children in vulnerable circumstances (Gladaxe system). • In France, intelligence services deployed algorithms that detect anomalous behaviour from internet users. • In Spain, an algorithm decides if tenants are eligible for subsidised electricity prices using income and rent data. • Polish tax authorities use STIR, an algorithm that sifts through the data of millions of entrepreneurs in order to fight tax fraud. The system can automatically block entrepreneurs' accounts based on the result of STIR analysis.	High risk due to 'black box' problems (i.e. lack of transparency and/or predictability in the inner working of the algorithms used); Long-term impact on eroding privacy because many devices and entities gather data without users' full understanding.	G2C, less B2B	Data are available and (involuntarily) used.	Little to no intervention if scoring is automated.
ADM in public administration • In Serbia, the e-Inspector software conducts risk assessments in trade, labour, administrative and sanitary areas. An algorithm sorts objects by risk levels based on static and dynamic data to plan inspections. The simultaneously implemented legislation prevents the inspector from inspecting an object classified as non-risk, thereby removing human oversight and control over the process.	High risk due to 'black box' problems (i.e. lack of transparency and/or predictability in the inner working of the algorithms used); Long-term impact on eroding privacy.	G2C	Data provision needed for application.	Little to no intervention because ADM process is automated.
Visa/residence permission • EU nationals apply to the UK 'Settled Status' programme through a mobile app for EU citizens to obtain permission to stay. Automated checks of previous records calculate whether a person is eligible. • UK visa applications are filtered by AI, affecting immigration policy decisions over who is allowed to enter the country. • The 'streaming tool' was an algorithmic system designed to categorise UK visa applications with reference to how much scrutiny each application needed. It would assign an application a green/amber/red rating. Red ratings meant the case worker ought to spend more time applying scrutiny and would have to justify approving the application to a more senior officer. Applications with a red rating were much less likely to be successful than those rated green, with around 99.5% of green being successful but only 48.59% of red. • In Canada, dozens of African researchers were denied visas for a leading AI conference.	High risk due to 'black box' problems (i.e. lack of transparency and/or predictability in the inner working of the algorithms used); Long-term impact on eroding privacy.	G2C	Data provision needed for application.	Little to no intervention because scoring is automated.
VI JUSTICE EU Charter of Fundamental Rights Articles 47-50. In particular: Right to an effective remedy and to a fair trial, Presumption of innocence and right of defence, Principles of legality and proportionality of criminal offences and penalties				
Biometrics and facial recognition in policing • In the US, the COMPAS system predicted higher risk values for black (and lower for white) defendants than their actual risk. • A US NIST study found that Asian and African American people were up to 100 times more likely to be misidentified than white men, depending on the particular algorithm and type of search, as well as high one-to-one error rates, one	High (potential to create distrust in public authorities).	G2C, less B2C	Data are collected (mostly involuntarily), databases are evaluated.	No means of intervention if monitoring and data collection methods (e.g. surveillance cameras) are not accessible to users. No means of redress if police are only authority involved.

of the most frequently used AI techniques in law enforcement. • Prison technology companies conduct voice recognition analysis on calls to generate unique voice prints. An algorithm uses recorded phrases of prisoners and stores the voice prints in a database, along with those of people found innocent. • In Austria, criminal police use a commercial, proprietary tool for automated face recognition. Key details, such as the tool's accuracy or the database of pictures to which it has access, are unknown. • Ugandan police work with Huawei to implement a 'safe city' surveillance system in the country. The installation is about 85% complete in the capital city of Kampala.				
Predictive policing • At least 11 local police forces in the EU automatically analyse images from surveillance cameras. Computer vision and facial recognition are linked to automated systems that claim to detect suspicious movements, such as driving in bus lanes, theft, assault or gatherings of aggressive groups. • In the Netherlands, the Dutch Crime Anticipation System (CAS) predicts more at-risk areas in a city in order to improve efficient distribution of their workforce. • In Lower Saxony, Germany, the PreMap project aims to predict domestic burglary based on historic crime data. • In the UK, the Harm Assessment Reduction Tool (HART) creates profiles for entry into diversion programmes on the basis of sensitive and personal information. The machine learning algorithm claims to assess a suspect's risk of reoffending, using over 30 variables, including criminal history and socio-demographic background data. • In the Netherlands, the 'ProKid' AI-tool aims to identify the risk of recidivism among 12-year old children previously suspected of a criminal offence. • In Chicago, Illinois, an algorithm rates every person arrested with a numerical threat score from 1 to 500-plus. Almost 400,000 Chicago citizens now have an official police risk score. The Strategic Subject List is based on an algorithm that Chicago police use to predict how likely it is that an individual will be involved in a shooting in the near future, as either shooter or victim. • By 2016, around 150 US police authorities tested predictive policing without scientific confirmation of the effectiveness of the systems.	High long-term risks and impacts to the right to a fair trial, non-discrimination and to create distrust in public authorities.	G2C, less B2B	Data are mostly gathered from various sources; citizens have little or no means to object and/or challenge the systems' profiling mechanisms.	No means of intervention if monitoring and data collection methods (e.g. surveillance cameras) are not accessible to users. No means of redress if police are only authority involved.

ANNEX 2: METHODOLOGY FOR ANALYSIS OF SUBMISSIONS TO THE PUBLIC CONSULTATION

In order to produce a timely and meaningful analysis of 408 position papers, this report undertook two key steps.

1. **Analysis of each position paper with a standardised template**

A standardised Excel template was created and completed for each position paper by a group of analysts. Each position paper was represented by one row and the columns were filled in with three main types of data:

(a) Multiple-choice questions: analysts answered closed multiple-choice questions, such as: 'What is the position paper's overall perception of the White Paper on AI?': 'Broadly positive'; 'Broadly negative'; 'Unclear'. For each closed question, analysts could add additional noteworthy comments to enrich the detail.

(b) Main arguments: analysts extracted up to three main arguments from each position paper.

(c) Summary: analysts summarised the position papers in four sentences or less.

The result was a raw Excel file with 408 rows and a rich set of datapoints for each position paper.

2. **Analysis of raw data**

The raw dataset was then cleaned, analysed and interpreted in Excel. The results have been transposed into the different sections in this report - each section contains key findings as well as more detailed tables and graphs.

All numbers should be read with an 'at least' qualifier ('at least 74 stakeholders believe that ...'), as a maximum of three main arguments were recorded for each paper. More stakeholders may share that position, but it may not have been one of their three main arguments. Similarly, for the other sections, only the positions that were explicitly mentioned were recorded – again, more stakeholders may hold that position but might not have made it explicit.

3. **Note on the number of position papers analysed**

A total of 422 respondents chose to submit position papers to the open public consultation. The research team also received 13 additional position papers from the Commission in the week before the deadline, which were included in the aggregate analysis. Position papers that the study team received after the deadline were analysed and summarised for the raw data tables but excluded from the aggregate analysis. Several respondents uploaded more than one document, for example academic papers unrelated to the White Paper (462 pdfs were uploaded). In those cases, only the position papers that directly targeted the AI White Paper were analysed. Some respondents uploaded duplicate documents, attached copies of the questionnaire, or less meaningful documents like flyers - these documents were not taken into account. As a result, 408 position papers were analysed.

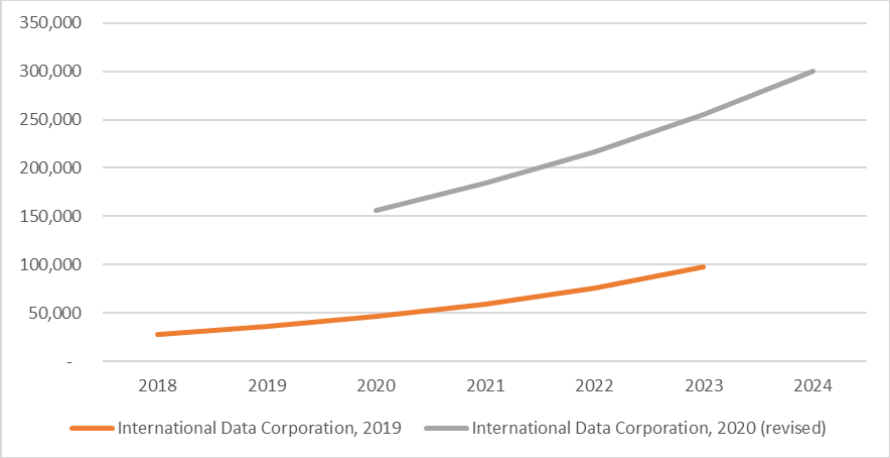
ANNEX 3: METHODOLOGY FOR ESTIMATION OF AI MARKET SIZE AND EVOLUTION

The team used a series of available estimates on the size and evolution of the AI market globally[90]. Although reported amounts are difficult to compare directly (analysts may use different definitions of AI), they are nevertheless useful as a guide. The European share of the global AI market was then assumed to be 22%, based on its share in AI software published in 2019 (Statista, 2019).

In July 2020, Grand View Research published the highest estimate for the AI market at the time. This was the first post-pandemic estimate available. The report summary refers to the coronavirus pandemic as 'an opportunity for AI-enabled computer systems to fight against the epidemic, as several tech companies are working on prevent, mitigate, and contain the virus' (Grand View Research, 2020).

In September 2019, International Data Corporation (IDC) predicted global AI market growth from USD 37.5 billion in 2019, to USD 97.9 billion in 2023 (IDC, 2019). IDC later revised that estimate to USD 156.5 billion in 2020, to eventually exceed USD 300 billion by 2024 (IDC, 2020).

Figure 25 - IDC estimates of global AI market (USD million)



Source: Visualisation of IDC data by authors

This significant revision upwards suggests that previous estimates will be less reliable.

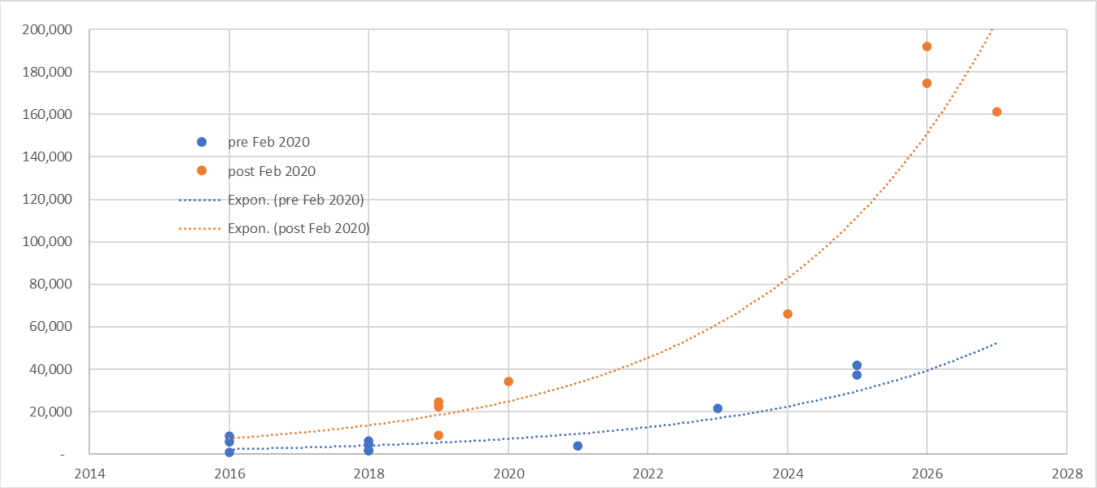
In September 2020, Zion Market Research provided a sample report on the European AI market, which was estimated to grow from USC 22.5 billion in 2019 to USD 174.5 billion in 2026. This is in line with the IDC estimate, assuming a European share of the global AI market of 22%.

Conclusion

Forecasts made after the COVID-19 pandemic are significantly higher, enabling the use of two types of forecasts, those published before February 2020 (pre-Feb 2020) and those after (post-Feb 2020). One of each estimate was used for a higher and lower-bound. The 'high growth' scenario is believed to be more likely, given the agreement between the most recent estimates, which account for the latest developments, such as a push to digitisation due to pandemic-related movement restrictions. The lower bound is used as a precaution against the event of a 'digital bubble'.

Figure 26 - European AI market size (USD million)

[90] ReportLinker, OECD (based om Crunchbase), CB Insights, McKinsey Global Institute, Grand View Research, Allied Market Research, Statista/Tractica, OMDIA/Tractica, UBS, Markets and Markets, McKinsey, IDC, Zion Market Research.



Source: Authors elaboration based on various market research estimates

From the existing estimates, exponential growth is the most likely scenario, backed by Grand View Research (2020). The rate of exponential growth was calculated using the initial and final values for the forecast period, while the average compound annual growth rate (CAGR)[91] was deduced and the years in between estimated.

[91] CAGR=(Market value in the final time periodMarket value in the initial time period)1/(number of time periods)

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

– at the following standard number: +32 22999696 or

– by email via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

Publications Office
of the European Union