AI4PEOPLE'S 7 AI GLOBAL FRAMEWORKS

AI IS NOT MERELY
ANOTHER UTILITY THAT NEEDS
TO BE REGULATED
ONLY ONCE IT IS MATURE.

IT IS A POWERFUL FORCE
THAT IS RESHAPING
OUR LIVES, OUR INTERACTIONS,
AND OUR ENVIRONMENTS.

Luciano Floridi
2018 Chairman, Scientific Committee AI4People, Professor of Philosophy and Ethics of Information and Director of the Digital Ethics Lab at Oxford University.

AI4PEOPLE'S 7 AI GLOBAL FRAMEWORKS

Following its past work on AI ethics (with the "*AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*") and on AI governance (with the "*AI4People Report on Good AI Governance: 14 Priority Actions, a S.M.A.R.T. Model of Governance, and a Regulatory Toolbox*"), in 2020 AI4People has identified seven strategic sectors (Automotive, Banking & Finance, Energy, Healthcare, Insurance, Legal Service Industry, Media & Technology) for the deployment of ethical AI, appointing 7 different committees to analyze how can trustworthy AI be implemented in these sectors: the *AI4People's 7 AI Global Frameworks* are the result of this effort.

TABLE OF CONTENTS

TABLE OF CONTENTS

COMMITTEES MEMBERS

Atomium EISMD wishes to thank the following Chairs and Committee's members for their participation and contributions:

Automotive: Christoph Lütge[1]; Franziska Poszler[2]; Aida Joaquin Acosta[3]; David Danks[4]; Gail Gottehrer[5]; Nicolae Lucian Mihet[6]; Aisha Naseer[7]; **Banking & Finance**: Nir Vulkan[8]; Aisha Naseer[7]; Frank McGroarty[9]; Giulia Del Gamba[10]; John Cooke[11]; Lampros Stergioulas[12]; Paul Jorion[13]; Raffaella Donini[14]; **Energy**: Nicolae Lucian Mihet[6]; Afzal S. Siddiqui[15]; Fausto Pedro García Márquez[16]; Rónán Kennedy[17]; Sergio Saponara[18]; **Healthcare**: Raja Chatila[19]; Stephen Cory Robinson[20]; Donald Combs[21]; Paula Boddington[22]; Hervé Chneiweiss[23]; Eugenio Guglielmelli[24]; Danny van Roijen[25]; Jos Dumortier[26]; Leonardo Calini[27]; **Insurance**: Frank McGroarty[9]; Gianvito Lanzolla[28]; Nir Vulkan[8]; Paul Jorion[13]; Patrice Chazerand[29]; Rui Manuel Melo Da Silva Ferreira[30]; Tilman Hengevoss[31]; Xenia Ziouvelou[32]; **Legal Services Industry**: Burkhard Schafer[33]; Cornelia Kutterer[34]; Elisabeth Staudegger[35]; Evdoxia Nerantzi[36]; Jacob Slosser[37]; Jamie J. Baker[38]; Mireille Hildebrandt[39]; Rónán Kennedy[17]; **Media & Technology**: Jo Pierson[40]; Stephen Cory Robinson[20]; Paula Boddington[22]; Patrice Chazerand[29]; Aphra Kerr[41]; Stefania Milan[42]; Fons Verbeek[43]; Cornelia Kutterer[34]; Evdoxia Nerantzi[36]; Elizabeth Crossick[44]; Norberto Andrade[45]; Janne Elvelid[46].

1. **Chairman Automotive Committee, AI4People; Director of the TUM Institute for Ethics in Artificial Intelligence at Technical University of Munich, Germany**
2. **Research Associate & PhD Student, Technical University of Munich, Germany**
3. **Head of Unit at Ministry of Transport and Infrastructure, Madrid, Spain**
4. **L.L. Thurstone Professor of Philosophy and Psychology Chief Ethicist, Block Center for Technology and Society, Carnegie Mellon University, USA**
5. **Law Office of Gail Gottehrer LLC A Law Firm Focused on Emerging Technologies**
6. **Chairman Energy Committee, AI4People; Professor in Energy Technology, Faculty of Engineering, Oestfold University College, Norway**
7. **AI Ethics Research Manager at Fujitsu Laboratories of Europe**
8. **Chairman Banking & Finance Committee, AI4People; Associate Professor of Business Economics at Saïd Business School, University of Oxford, UK**
9. **Chairman Insurance Committee, AI4People; Professor of Computational Finance and Investment Analytics; Director of Centre for Digital Finance at Southampton Business School, UK**
10. **Digital and Innovation Policy Advisor at Intesa Sanpaolo**
11. **Chairman of the Liberalisation of Trade in Services Committee at TheCityUK**
12. **Professor in Business Analytics at the University of Surrey, UK**
13. **Associate Professor of Ethics, Université Catholique de Lille, France**
14. **Senior Manager, European Digital and Innovation policies at Intesa Sanpaolo**
15. **Professor of Energy Economics in the Department of Statistical Science, UCL, UK**
16. **Full Professor at Castilla-La Mancha University, Spain**
17. **Lecturer in Law, School of Law, National University of Ireland Galway, Irelands**
18. **Professor of Electronics, Department of Information Engineering, Pisa University, Italy**
19. **Chairman Healthcare Committee, AI4People; Professor and Director of the Institute of Intelligent Systems and Robotics (ISIR) at Pierre and Marie Curie University in Paris, France**
20. **Senior Lecturer/Assistant Professor in Communication Design at Linköping University, Norrkoping, Sweden**
21. **Vice President & Dean of the School of Health Professions, Eastern Virginia Medical School, USA**
22. **Senior Research Fellow, New College of the Humanities London, UK**
23. **Directeur de Recherche au CNRS, Paris, France**
24. **Senior Advisor on Publications for IEEE RAS Professor of Bioengineering Prorector for Research Founder, Research Unit of Biomedical Robotics and Biomicrosystems Università Campus Bio-Medico di Roma**
25. **Digital Health Director at COCIR**
26. **Honorary Professor of ICT Law at the University of Leuven, Belgium**
27. **Policy Manager, European Government Affairs at Microsoft**
28. **Professor and Dean at Cass Business School - City, University of London, UK**
29. **Director at DIGITALEUROPE**
30. **Chief Data Governance Officer, Zurich Insurance Group (ZIG)**
31. **Head Public Affairs EMEA Region at Zurich Insurance Group (ZIG)**
32. **Innovation Officer and Research Scientist, Institute of Informatics and Telecommunications, National National Centre for Scientific Research Demokritos, & Member of the Scientific Committee on Data Policy and Artificial Intelligence, National Council for Research and Innovation (NCRI), Greece**
33. **Chairman Legal Services Industry Committee, AI4People; Professor of Computational Legal Theory; Director, SCRIPT Centre for IT and IP Law, University of Edinburgh, Scotland**
34. **Senior Director, Rule of Law & Responsible Tech, European Government Affairs at Microsoft**
35. **Professor at Universität Graz, Austria**
36. **Policy Manager, European Government Affairs at Microsoft**
37. **Carlsberg Foundation Postdoctoral Fellow at University of Copenhagen, Denmark**
38. **Associate Dean and Director of the Law Library; Professor of Law at Texas Tech University School of Law, USA**
39. **Research Professor on 'Interfacing Law and Technology' at Vrije Universiteit Brussel, Belgium**
40. **Chairman Media & Technology Committee, AI4People; Professor at imec-SMIT, Department of Media & Communication Studies, Vrije Universiteit Brussel, Belgium**
41. **Professor of Sociology at Maynooth University and Maynooth lead of the ADAPT Centre for Digital Media Technology, Ireland**
42. **Associate Professor of New Media and Digital Culture, University of Amsterdam**
43. **Full Professor in Bio-Imaging and Bio-Informatics, Leiden Insitute of Advanced Computer Science**
44. **Head of Government Relations at RELX**
45. **Global Policy Lead for Digital and AI Ethics at Facebook**
46. **Policy Manager EU Affairs at Facebook**

AI4P

IN BRIEF

AI4People is a multi-stakeholder forum, bringing together all actors interested in shaping the social impact of new applications of AI, including the European Commission, the European Parliament, civil society organisations, industry and the media.

Launched in February 2018 with a three year roadmap, the goal of AI4People is to create a common public space for laying out the founding principles, policies and practices on which to build a "good AI society". For this to succeed we need to agree on how best to nurture human dignity, foster human flourishing and take care of a better world. It is not just a matter of legal acceptability, it is really a matter of ethical preferability.

1

AUTOMOTIVE

AI4People-Ethical guidelines for the automotive sector: Fundamental requirements & practical recommendations for industry and policymakers

Authors

Christoph Lütge
Director of the TUM Institute for Ethics in Artificial Intelligence at Technical University of Munich, Germany

Franziska Poszler
Research Associate & PhD Student, Technical University of Munich, Germany

Aida Joaquin Acosta
Head of Unit at Ministry of Transport and Infrastructure, Madrid, Spain

David Danks
L.L. Thurstone Professor of Philosophy and Psychology Chief Ethicist, Block Center for Technology and Society, Carnegie Mellon University, USA

Gail Gottehrer
Law Office of Gail Gottehrer LLC A Law Firm Focused on Emerging Technologies

Nicolae Lucian Mihet
Professor in Energy Technology, Faculty of Engineering, Oestfold University College, Norway

Aisha Naseer
AI Ethics Research Manager at Fujitsu Laboratories of Europe

ABSTRACT

This paper presents the work of the AI4People-Automotive Committee established to advise more concretely on specific ethical issues that arise from autonomous vehicles (AVs). Practical recommendations for the automotive sector are provided across the topic areas: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing as well as accountability. By doing so, this paper distinguishes between policy recommendations that aim to assist policymakers in setting acceptable standards and industry recommendations that formulate guidelines for companies across their value chain. In the future, the automotive sector may rely on these recommendations to determine relevant next steps and to ensure that AVs comply with ethical principles.

Keywords: Autonomous driving, Self-driving cars, Autonomous vehicle ethics, Governance, Regulation, Ethics of AI, AI4People, Transparency, Override, Fundamental Rights

AIM & SCOPE OF THIS PAPER

In the past decade, many policy documents have discussed ethical issues and potential future directions related to new emerging technologies such as artificial intelligence (AI) or autonomous systems. This paper presents the work of the AI4People-Automotive Committee[1] established to advise more concretely on specific ethical issues that arise from autonomous vehicles (AVs). The committee consisted of industry experts and researchers from the fields of ethics, law, philosophy, engineering, technology and policy. The aim of this paper is to provide the automotive sector, including both companies and public entities such as regulators, with concrete and practical guidelines to comply with ethical principles within the AI systems of AVs. Therefore, this paper could serve as a checklist for policymakers and companies as well as a basis for developing a certification of ethics, an 'ecosystem of trust' (European Commission, 2020b) and ultimately a 'Good AI Society' (Floridi et al., 2018) in the automotive sector. These guidelines are intended to provide a clearer vision and moral compass on how to proceed and what to consider when developing AVs, rather than additional barriers to innovation. The automotive sector is defined here in the broadest terms possible to encompass a wide range of companies involved in the development of vehicles, including private cars, trucks, busses and shuttles. Sea, air and military-type applications have been excluded due to their functional and ethical specificity. This paper will focus on the ethics of the AI-based tools that are used in automotive technology, rather than on the ethics of vehicles in general.

1 All co-authors of this paper constitute the AI4People-Automotive Committee.

This paper distinguishes between high-level guidelines for policymakers ('policy recommendations') and concrete actionable recommendations for companies ('industry recommendations'). However, the line between the two cannot always be drawn clearly which also highlights the importance of co-regulation (i.e. the interaction of legal regulation and self-regulation by companies) (Pagallo et al., 2019). The policy recommendations are designed to focus attention on pressing policy issues and assist in setting acceptable standards. Thus, the policy recommendations ultimately influence the industry recommendations. Responsible targets for the execution of the policy recommendations are: policymakers, legislators, ethics standards boards and commissions such as the United Nations Economic Commission for Europe (UNECE). The industry recommendations formulate guidelines for companies across their entire value chain (especially during research & development, production & operations and service). Therefore, original equipment manufacturer (OEM) / car manufacturers are the primary responsible targets for those recommendations.
Before turning to the principles and guidelines, we note three points of consensus among the authors: (1) a responsible balancing of risks or estimated harm should be permitted at any time for AVs; (2) a large-scale introduction of full-mode AVs (level 4 and higher) onto streets is unlikely in the short run, so we must consider a more incremental, step-by-step approach; and (3) policymakers face significant challenges now, and so there are significant pressures to quickly develop a clear regulatory framework.

THE GUIDELINES

Fundamental rights underlying the guidelines

Particular fundamental rights are the basis for the proposed seven requirements that were originally derived by the High-level Expert Group on Artificial Intelligence (2019) (i.e. human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing; accountability) and recommendations in this paper. In addition to general human dignity, key fundamental rights (United Nations, 1948) that policymakers and companies in the automotive sector should recognize are: *Right to self-determination and liberty* which draws attention to human agency in self-driving cars (i.e. importance of override options) (see Guideline 1). *Right to life and security* which entails ensuring technical robustness and safety of operating self-driving vehicles; on a broader

level, this includes securing societal and environmental wellbeing (see Guideline 2 and 6). *Right to protection of personal data* drawing attention to data ownership, data governance and privacy of personal data that is generated during the operation of self-driving cars (see Guideline 3). *Right to equality and non-discrimination* requiring the avoidance of unfair bias in operating vehicles as well as the accessibility of benefits for every individual in society (see Guideline 5). *Right to explanation* which, in the field of autonomous driving, demands transparency and communication of the underlying functionality, which can be achieved through accountability measures such as audits and logging mechanisms (see Guideline 3 and 7). Certainly, incompatibilities and trade-offs between fundamental rights can emerge; for example, life and security can be in tension with the right to self-determination. On the one hand, AVs are expected to improve traffic flow and decrease fatalities that are due to human error. On the other hand, automated driving systems reduce the driver's autonomy, perhaps to the point of being a mere passenger. In this regard, the Ethics Commission on Automated and Connected Driving (BMVI, 2017) formulated the following guideline: "In a free society, the way in which technology is statutorily fleshed out is such that a balance is struck between maximum personal freedom of choice in a general regime of development and the freedom of others and their safety" (Lütge, 2017, p. 550). In conflict situations, policymakers and legislators should decide which fundamental rights are to be prioritized.

Policy recommendations:

- Relevant fundamental rights to be considered in the field of autonomous driving are: human dignity, right to self-determination and liberty, right to life and security, right to protection of personal data, right to equality and non-discrimination as well as the right to explanation.
- It must be realized that there will be no technologies or policies that maximize all fundamental rights for everybody simultaneously. There will always be trade-offs. Therefore, policymakers and legislators should decide which fundamental rights are to be prioritized in particular situations.
- In doing so, policymakers and legislators should cooperate with multiple stakeholders to obtain necessary information for executing an evaluation and subsequent agreement on compromises and prioritization.

1. Human agency and oversight – including monitoring, training, human-machine interfaces and external control of vehicle data

A few guidelines have already been developed that highlight the importance of maintaining personal autonomy in AVs, including possible requirements for a 'stop' or 'override' button (European Commission, 2020b; Lütge, 2017). At the same time, autonomy requires informed and deliberate control, and so overrides (and other measures) should not necessarily be universal. In particular, admissibility of human **override should be conditional** on two aspects:

(1) The level of automation of the AV[2]

- for levels up to and including 3, there should be an override function that can be executed at any time.
- for level 4, there should be an override function that can be executed only when not impacting or undermining the safety mechanisms of the AV (e.g., one helpful factor to satisfy this requirement might be to implement overrides with a time lag). The rationale for this is that, if individuals were allowed to intervene immediately at any point, the inherent logic and longer-term plan completion of the technically functional AV is disrupted which may lead to increased risks for all parties involved.
- for level 5, it is not necessary to include an override function, as it would take away many of the original advantages such as inclusive accessibility (e.g., by excluding elderly, disabled individuals, youth or individuals who do not hold a driving license), safety (e.g., humans taking control may be out of practice), trust (e.g., giving drivers the impression that the system could fail), and comfort (e.g., limiting opportunities for new and more comfortable mobility options and designs)[3]

(2) The state and behavior of the driver

- when the driver's abilities are impaired (e.g., due to alcohol consumption), the availability of an override function should be limited and preceded by a request for confirmation

Nevertheless, recent examples of AVs involved in crashes draw attention to the failing assumption of responsibility by individuals. The underlying problems relate to overconfidence in, or overreliance on, the AI system as companies do not adequately warn drivers and/or drivers violate the guidelines provided by the companies.

2 The levels refer to the taxonomy developed by the SAE International (2018) for six levels of driving automation, ranging from no driving automation (level 0) to full driving automation (level 5).
3 The override function does not need to be similar to the way we are driving today such as taking over using a steering wheel or a paddle. On the contrary, the control can be a function provided through some interfaces that do not take away the original advantages of AVs such as inclusive accessibility.

Tesla Highway Accident: In 2018, a Tesla's Model X car crashed into a curb, collided with two other vehicles and caught fire while in Autopilot mode. The Tesla driver died from blunt-force trauma injuries. The U.S. National Transportation Safety Board (2020) determined that the probable cause of the crash was related to system limitations of the Tesla Autopilot, as well as the driver's overreliance on the system and lack of response (due to distraction likely from a cell phone game application). Tesla's position was that it tries to ensure and monitor driver engagement in order to prevent driver overreliance, and that its policies advise Tesla owners that in an SAE-defined Level 2 partial driving automation system (which it considers its vehicles to be), it is the driver's responsibility to be prepared to intervene at all times. Nevertheless, drivers continue to be overly reliant on Autopilot and appear to believe that when in Autopilot, the vehicle is fully autonomous. This raises questions about the effectiveness of Tesla's disclosures of the capabilities of the vehicle when in Autopilot. This case highlights that appropriate agency requirements must go beyond giving the driver the option to use a stop-button and include providing the driver with sufficient information and training to know when to press that button.

Therefore, companies must clearly distinguish and make apparent whether a driverless system is being used or whether a driver remains accountable for driving (Lütge, 2017). In order to realize effective human agency and clarity over personal responsibility, our approach concerning AVs is threefold:

Monitoring, training and an external human-machine interaction is needed to improve one's ability to act with intention.

1. Companies should put in place technical safeguards to help drivers remain fully aware and ready to take over the driving when the AV expects them to. AVs should **monitor drivers** and help drivers remain awake and attentive. For example, current driving monitoring systems using camera-based facial recognition technology determine the driver's level of vigilance and trigger alerts to the driver when signs of distraction are detected (Research & Markets, 2019). Other monitoring systems are related to the amount of torque in the steering wheel. For example, Tesla (2020) locks the activation of the autopilot mode if the driver seems inattentive (e.g., insufficient torque is applied or warnings are repeatedly ignored). The upcoming regulation on automated lane keeping systems will obligate car manufacturers to introduce driver availability recognition systems and clarify the criteria that assess whether a driver is deemed to be unavailable (e.g., eye closure) (UNECE, 2020b). UNECE also considers that "[a]utomated/autonomous vehicles should include driver engagement monitoring in cases where drivers could be

involved (e.g., take-over requests)" (UNECE, 2019, p. 3). It is important that handovers be aligned with the level of automation: As the level of automation increases, drivers engage more in other activities such as watching a video, which decreases human capability to take over control (Merat et al., 2014). Thus, handovers should conform to human capabilities by, for example, obviating "the need for an abrupt handover of control to the driver (emergency)" (Lütge, 2017, p. 556). There is currently no agreement on what constitutes a comfortable transition time, and so we do not propose a universal prescription on this point. In the meantime, companies should provide documentation that justifies their particular handover window. A possible starting point for determining a reasonable transition time might be that AVs, as they drive, could learn about the capabilities of drivers from aggregated traffic data and adjust the vehicle's parameters accordingly (respecting a safe minimum time response).

2. Companies should **train drivers** on the capabilities and limitations of AVs (European Commission, 2020b), so that individuals can make informed decisions and do not over rely on the vehicle's capabilities (see also UNECE, 2019). This training should be tailored to different demographic groups, given recent studies that show demographic differences in interactions with AVs (Manser et al., 2019). Training programs should cover topics such as the "[system's] functional intent, operational parameters, system capabilities and limitations, engagement/disengagement methods, HMI, emergency fallback scenarios, operational design domain parameters (i.e., limitations), and mechanisms that could alter [the system's] behavior while in service" (NHTSA, 2017, p. 15). Drivers should also be trained on the purpose of using an AVs, the degree of automation, and conditions for potential system failures (Manser et al., 2019).
3. The importance of human autonomy applies not only to drivers but also to humans outside the vehicle such as pedestrians. Therefore, companies should ensure that these latter individuals can also exercise their autonomy. For example, AVs should have mechanisms to show pedestrians that they have been recognized and reveal the AV's motion intentions, perhaps with LED strips to convey perception information (e.g., displaying cool colors for far away obstacles and warm colors for near obstacles in the environment) (Florentine et al., 2016). These **external human-machine interfaces** facilitate human agency for pedestrians, as they enable them to feel less anxious about the technology and have more information to move freely and safely. However, further research is required to determine the most useful interfaces (Rouchitsas & Alm, 2019).

Additionally, **external oversight mechanisms** need to be put in place to control for adequate human agency. Therefore, although internal overriding functions may not always or immediately be available for (drivers in) AVs, general oversight should be possible at all times. Live and total oversight is both impracticable and unwarranted

(Lütge, 2017). However, under certain circumstances, such as following a fatal accident, and depending on the legal and regulatory framework in place in the country where the accident occurred, it may be appropriate to designate an organization in each jurisdiction that is permitted to retrospectively look at the code and data within the AV to determine the cause of the accident (for more information see Guideline 7).

Industry recommendations:

- There should be a conditional override option allowing the control to be handed back to the driver. The admissibility of an override function depends on the level of automation of the AV (up to level 3: at any time; level 4: corresponding to safety mechanisms of an AV; level 5: not required) as well as on the state and behavior of the driver (e.g., impaired ability).
- AVs should continuously assess and monitor the driver's attentiveness and ability to intervene. Before operation, the AV could pose control questions to the driver (e.g., did you ingest any drugs or alcohol?); during operation, the AV could use sensors and biometric technology to do so. The upcoming UN Regulation on Automated Lane Keeping Systems can serve as a baseline for car manufacturers to develop appropriate driver attentiveness recognition systems.
- Handover should correspond to the driver's capabilities. Therefore, AVs could learn about drivers' capabilities and response times during operation from aggregated data and adjust the vehicle's parameters accordingly (respecting a safe minimum time response).
- Companies should provide documentation that justifies their particular handover window.
- Training programs should be tailored to different demographic groups and exhibit minimum elements that should be regarded in a training curriculum (e.g., limitations and capabilities of AVs) based on findings of recent studies.
- AVs should offer a 'training mode', for the first kilometers to train drivers on the AV's functioning.
- External human-machine interfaces should clearly communicate about the vehicle's motion intention and awareness of other traffic participants to humans outside the vehicle.

Policy recommendations:

- Policymakers should finalize what constitutes acceptable and legitimate override functions and define applicable situations for activation.
- Policymakers should determine standards for drivers' monitoring, training requirements, handover routines and external human-machine interfaces. These standards should be as global as possible.

- Policymakers in each jurisdiction should consider designating an organization in each jurisdiction that is allowed to look at the code and data within the AV in the event of a fatal accident involving an AV or a corresponding legal proceeding.

2. Technical robustness and safety – including resilience to attack and security, fall back plan and general safety, accuracy and reliability

A prime requirement of AVs should be safety, both in ordinary operations and if subject to adversarial attack (Lütge, 2017).

There are many **differing forms of potential threats** to AVs, and so governmental entities such as the ENISA (2019) or UNECE (2020a) have created holistic summaries and categorizations of relevant dangers and vulnerabilities. Firstly, there are threats that do not solely apply to AVs but also to conventional vehicles such as *technical malfunctions and outages* including sensor and other failures (ENISA, 2019). Secondly, there are threats that are particularly important for AVs and can be subsumed under the term 'cybersecurity'. Potential cybersecurity threats include the following:

- *hijacking* such as unauthorized information disclosure or extraction of copyrighted or proprietary software from vehicle systems (product piracy) (UNECE, 2020a)
- *abuse* such as attacks on back-end servers that stops the vehicle's functioning (e.g., disruptions of communication and external connectivity) or threats regarding the vehicle's update procedures (e.g., preventing the rollout of critical software updates) (UNECE, 2020a)
- *passive behavioral attacks* such as individuals intentionally interfering with AVs. For example, human drivers might tend to drive more aggressively around AVs or jaywalking may increase because it is known that AVs respect the safety distance.

There are several categorizations of threats that relate to the data stored in vehicles on an associated server and to the information exchanged during communication between the vehicle and the server. These threats can impact the safe operation of the vehicle, alter the software operation, and generate data breaches, though many of these threats are not specific to AVs but also can be found in current vehicles.

It is essential to develop mechanisms to **test an AV's cybersecurity management system** before operation. The EU Cybersecurity Act aims to establish a general certification framework for ICT digital products, services, and processes that

While oversight is more about retrospect, safety is more about prospect.

allows the "creation of tailored and risk-based EU certification schemes" (ECCG, 2020). Similarly, the UN is preparing a regulation on uniform provisions concerning the approval of vehicles with regard to cybersecurity and of their cybersecurity management systems. For example, the draft regulation (as of March 2020) proposes an international approval mark or the verification of a manufacturer's compliance by an approval authority (UNECE, 2020a). In the future, such clear regulations and standardized tests will be necessary so that all companies are informed about, and comply with, the universal requirements for cybersecurity management systems. Governments should "promote mutual recognition systems and certification schemes that are built upon international standards [...] to facilitate international harmonization on privacy and security" (Joaquin Acosta, 2019, p. 215). SAE J3061, a comprehensive cybersecurity implementation guideline for the automotive industry, can serve as a starting point (SAE International, 2016).

Eurocybcar – Cybersecurity test for cars: Vehicles can be considered computers on wheels. They contain systems such as ABS, airbags, Bluetooth, eCall and remote control keys which make the vehicle susceptible to cyberattacks. Therefore, Eurocybcar developed the first European testing program for verifying the level of cybersecurity of (autonomous) vehicles. The test is twofold: first, it assesses the level of protection against cyberattacks that a vehicle has; second, it evaluates how a cyberattack would affect the integrity of the car's system and the physical security and privacy of its passengers. As soon as a car passes the Eurocybcar test, it receives the 'Cybersecure Car' seal, with a rating of one to five (Eurocybcar, 2019).

Additional to threats, measures need to be developed that **assess the general functionality of an AV**. The Ethics Commission on Automated and Connected Driving suggests that "[t]he public sector is responsible for guaranteeing the safety of the automated and connected systems introduced and licensed in the public street environment. Driving systems thus need **official licensing** and monitoring" (Lütge, 2017, p. 550). For example, a kind of TÜV, i.e. a technical inspection agency, for AVs could be developed. Relevant factors to be assessed here are accuracy, reliability and fallback options of AVs. In terms of **accuracy and reliability**, it could be tested to what extent the AV's underlying "AI meets, or exceeds, the performance of a competent & careful human driver", refrains from engaging in "careless, dangerous or reckless driving behavior" as well as to what extent it "remains aware, willing and able to avoid collisions at all times" (ADA, 2020). SAE International published a more detailed and elaborate list of driving safety performance assessment metrics such as minimum safe distance factors or proper responses (Wishart et al., 2020). Furthermore, **safeguards against technical failures** and outages need to be established. The IEEE P7009

standard for fail-safe design of autonomous and semi-autonomous systems could serve as a baseline for developers. The standard provides clear procedures for measuring, testing, and certifying a system's ability to fail safely as well as instructions for improvement in the case of unsatisfactory performance (IEEE, 2019).

In terms of **general safety and fallback plan**s, "[i]n emergency situations, the vehicle must autonomously, i.e., without human assistance, enter into a '**safe condition**'" (Lütge, 2017, p. 556). This condition has been specified by proposing the terms 'minimal risk condition' and 'minimum risk maneuver'. The **Minimal risk condition** is "[a] condition to which a user or an ADS may bring a vehicle after performing the DDT [dynamic driving task] fallback in order to reduce the risk of a crash when a given trip cannot or should not be completed" (SAE International, 2018, p. 11). The "**Minimum risk maneuver** means a procedure aimed at minimizing risks in traffic, which is automatically performed by the system" (Leonhardt, 2018, p. 12). Causes for the execution of such a maneuver could be detection that the driver is inactive and not reacting to transition demands, or reaching system failure / boundaries when the driver is not responding to transition demands. In such situations, potential maneuvers could entail "further lane keeping for a certain time, enlarging gap to other road users, [...] slowing down to standstill" (BMVI, 2015, p. 4). What constitutes an appropriate maneuver depends on (1) the operation condition of the vehicle (e.g., technical failures that hinder the AV to perform a fallback), (2) the prevailing environmental conditions (e.g., density of traffic) and (3) regulatory boundary conditions (Leonhardt, 2018). Although SAE J3016 (Leonhardt, 2018) makes significant progress regarding the nature of a safe or minimal risk condition, the definition of such conditions as well as the particular circumstances in which such conditions should be activated (e.g., incidents that leave the driver incapacitated such as a stroke) need to be further determined and harmonized.

Overall, experimenting with new AVs and testing their technical robustness and safety should follow a stepwise approach: For example, "the levels of testing that should be conducted before **testing on open roads**, including, for example, the use of **simulation**, **hardware-in-the-loop testing**" should be identified and standardized (European Commission, 2020a, p. 29). Recognizing the challenges of physical test strategies for AVs (length of time they take to complete, high number of hours of drive time required), ESTECO has developed a white-box / scenario-based verification system to investigate the performance of ADAS/AD functions across different sensors, algorithms, actuation and scenarios (ESTECO, 2020). Systems like these can act as helpful antecedents to actual testing on open roads.

Industry recommendations:

- The prime requirement of AVs should be safety.
- In addition to threats that relate to conventional vehicles, manufacturers of AVs should particularly focus on cybersecurity threats. In doing so, companies need to comply with regulations for cybersecurity management systems. SAE J3061 could serve as a guideline to design cybersecurity into AVs throughout the entire development life cycle process.
- In terms of general functionality and safety, vehicles need to pass an official test that assures the system's accuracy, reliability and adequacy of its fallback options. The SAE Driving Safety Performance Assessment Metrics and the IEEE P7009 standard could serve as a baseline to design fail-safe mechanisms of autonomous and semi-autonomous systems.

Policy recommendations:

- Regulations need to be developed that reflect consensus on the method by which to grant approval to a vehicle's cybersecurity management system.
- Policymakers need to work with industry experts to develop a standardized test for the general functionality and safety of AVs to assure the system's accuracy, reliability and adequacy of its fallback options. This test could serve as a basis for the approval of AVs for sale to consumers.
- Policymakers need to collaborate with industry experts to determine and harmonize the definition of a 'safe condition' / 'minimal risk condition', the corresponding 'minimum risk maneuvers', and the circumstances in which such maneuvers should be executed. In doing so, SAE J3016 could serve as a baseline.

3. Privacy and data governance – including respect for privacy, transparency and communication, and access to data

Conventional vehicles collect data through event data recorders (that record technical information about a vehicle's operation involved in crashes) and on-board diagnostic information (to access information about driver behavior, emission measures or diagnose performance issues). With new technological options, connected vehicles and AVs will make transportation safer and more convenient. However, many features depend on the collection and processing of ever more data in order to function effectively (Future of Privacy Forum, 2017).

Therefore, it is essential to specify the type and scope of data that AVs are permitted to collect. Three **types of data** can be distinguished that warrant special attention:

- *geolocation data* (e.g., for activating route navigation), which could reveal the passenger's location and life habits of individuals (EDPB, 2020)
- *biometric data* (e.g., for user recognition or tracking of driver's attention), which could be used to enable unauthorized access to a vehicle and enable access to a driver's profile settings and preferences (EDPB, 2020). The collection of this type of data applies not only to drivers but also to individuals outside the vehicle such as pedestrians.
- *driver behavior data*, which could reveal unlawful behavior, including traffic violations such as speeding (EDPB, 2020)

Some of this data will be collected automatically, and some will require consent from the vehicle owner or driver in order to activate and use certain functions. Careful consideration needs to be given to the collection of data from inside the vehicle that relates to things other than the operation of the vehicle. Additionally, individual's rights should be considered at **group level** (e.g., drivers versus pedestrians) (European Commission, 2020a). For example, data (especially, biometric data) of external parties such as individuals walking on the street should warrant special protection. Although the European Commission (2020a) has additionally problematized data collection when AVs pass through **particular locations** such as private and non-public settings, we suggest that collecting data in private spaces should not in general be restricted in order to guarantee an AV's functionality. The focus should instead be on the mode of data collection and sharing.

Overall, services that collect and share data should comply with all applicable laws, and be accompanied by a **strict privacy and data governance policy** that includes, but is not limited to, the following (Future of Privacy Forum, 2017):

1. **Transparency and communication**: Manufacturers need to provide clear and concise privacy policies to the vehicle owners that describe data collection and use. These policies must be readily understood by vehicle owners. These policies could, for example, be displayed in the purchase agreement, user manual or in the interface of an app. This is also in line with the General Data Protection Regulation (GDPR) stating that controllers must, before personal data is obtained, provide the data subjects with information necessary to ensure transparent processing about the existence of automated decision-making.
2. **Affirmative and explicit consent**: The driver's educated and affirmative consent is required before certain sensitive data is collected or used. This requirement is particularly critical for marketing uses, or if the data will be shared with unaffiliated

third parties. This is in line with the guideline of the Ethics Commission on Automated and Connected Driving about the permissibility to use data that is generated by AVs for other business models, which states that lastly "[i]t is the vehicle keepers and vehicle users who decide whether their vehicle data that are generated are to be forwarded and used" (Lütge, 2017, p. 555). Additionally, even in the absence of laws requiring it, users should always have the right to opt-out or request that particular data not be collected, unless those data are critical for the AV safe system's operation.

3. **Limited and useful sharing with third parties**: There should only be limited circumstances where manufacturers are allowed to share a vehicle's data with external parties. Under appropriate conditions and with the appropriate safeguards, data that guarantees safe operation of the vehicle and other traffic participants as well as data that provides benefits to overall society and is of public interest should be shared. For example, AVs could provide information to the local department of transportation about a pothole on the road, so that infrastructure inspections and maintenance resources can be better allocated (in consideration of a fair and unbiased distribution of resources) and traffic information can be shared to improve traffic flow and promote safety. Accordingly, the European Commission has issued a regulation requiring public or private road operators and service providers to share and exchange relevant road safety-related traffic data such as the observation of a temporary slippery road or exceptional weather conditions (European Commission, 2013). Personally identifiable information must always be given the highest levels of protection. If data must be shared with third parties due to the above mentioned reasons, they should be **anonymized and deidentified before being transmitted** (EDPB, 2020). For example, the EU Data Task Force partnered with TomTom to improve road safety by sharing anonymized vehicle and infrastructure data between countries and manufacturers. For example, this will allow the detection of dangerous road conditions such as slippery roads and issue warnings to other traffic participants. "The EU Data Task Force (DTF) will use a decentralised data collaboration architecture to share vehicle-generated data [...]. The datasets will then be taken by TomTom, processed further, and delivered back to other vehicles and road authorities via its live Traffic services" (Europawire, 2019). In line with Article 3(c) of Directive 2010/40/EU, data and procedures for the provision of road safety-related traffic information should be free of charge to users (European Commission, 2013). However, past studies showed individuals can sometimes be identified using anonymized data (Techcrunch, 2006; Archie et al., 2018), and so companies must ensure that the shared data does not permit re-identification (e.g., by **minimizing collected data** or using **differential privacy techniques**).

The anonymization issue is a pivotal point to be highlighted, because it distinguishes privacy from surveillance.

4. **Compliance with pertinent data protection standards and regulations**: All data collection and processing obviously must respect relevant regulations (EDPB, 2020), such as the GDPR that applies to the processing of personal individual data, as well as the ePrivacy directive for information access on the terminal equipment of a user (EDPB, 2020; European Commission, 2020b). The IEEE P7002 standard specifies how to manage privacy issues for systems that collect personal data, e.g., by providing a guideline for a privacy impact assessment (IEEE, 2019).

Industry recommendations:

- Manufacturers should follow a strict privacy and data governance policy that includes transparency and communication to users, requesting affirmative consent and allowing limited sharing of data with third parties (including governments). In doing so, companies should comply with applicable standards and regulations such as the GDPR, the ePrivacy directive or the IEEE P7002.
- Before transmitting personal information from an AV to third parties, steps must be taken to ensure that it cannot be traced back to an individual.
- Manufacturers should implement data protocols defining who can have access to data under which conditions.

Policy recommendations:

- Before receiving AV data, policymakers need to make clear what types of AV data they are seeking and how that data will enable them to improve public safety or some other legitimate public purpose (e.g., improve infrastructure, traffic flow and law compliance).
- At the EU level, building on Article 3(c) of Directive 2010/40/EU, consideration should be given to expanding the list of events and relevant traffic information that should be communicated free of charge.

4. Transparency – as a key mechanism to realize all other requirements

Transparency is more like a mean to an end – it is a key mechanism to realize the other six requirements.

In the automotive sector, we contend that transparency is not a freestanding desideratum, but rather a key mechanism to realize the other principles or requirements. Transparency plays a major role for achieving the principle of privacy and data governance, requiring that manufacturers provide vehicle owners with information regarding data collection practices and intended uses (for more information see Guideline 3). Similarly, to satisfy the principle of accountability, the implementation of explicit transparency measures such as logging mechanisms or black boxes are essential (for more information see Guideline 7). The IEEE P7001 ("Transparency of Autonomous Systems") standard can serve as a baseline to address these issues.

5. Diversity, non-discrimination and fairness – including the avoidance of unfair bias, responsible balancing and accessibility

Generally and regarding the operations of AVs, **no distinction between individuals** should be allowed and fair treatment of all humans should be enacted. This is clearly stated in the Universal Declaration of Human Rights: "[e]veryone is entitled to all the rights […] without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status" (United Nations, 1948, p. 2; Kriebitz & Lütge, 2020). In the field of AI (e.g., AVs), this obligation becomes ever more important as implicit biases and discrimination may unintentionally, and without transparency, be incorporated into algorithms. Studies show that systems can have differential performance for people of different ethnic groups, which consequently can result in them being harmed. For example, a study from the Georgia Institute of Technology illustrates how state-of-the-art AI object detection systems are less likely to detect pedestrians with darker skin color than those with lighter skin (Wilson, Hoffman & Morgenstern, 2019). Another study from the US National Institute for Transportation and Communities investigated the driving behavior through crosswalks that "revealed that black pedestrians were passed by twice as many cars and experienced wait times that were 32% longer than white pedestrians". If such driving data is fed into a machine-learning algorithm, the system may discover this discriminatory pattern and adapt it into its functioning (Forbes, 2020).

In order to ensure non-discriminatory programming and functioning, the systems need to be **trained and tested for unfair bias**. Companies should test their algorithms for bias and discrimination and demonstrate that certain fairness standards are met (Vox, 2019). Laws could be enacted, for example, that mandate that facial recognition software used by public entities and companies must be **independently tested** for bias (Secretary of State Washington, 2020). The IEEE P7003 standard for algorithmic bias considerations sets out instructions for eliminating bias when developing algorithms: it provides developers of algorithms for autonomous systems with protocols and includes criteria for selecting validation data sets (IEEE, 2019). The **training should be different depending on the location** where the system operates: when a technology is launched into the market, companies could localize it using location specific data. Companies could ensure that their **development teams are sufficiently diverse** to guard against intentional and implicit bias being incorporated into their algorithms and technologies (Vox, 2019).

The Moral Machine Experiment: The Moral Machine Experiment by Awad et al. (2018) is an online experimental platform designed to explore the moral dilemmas faced by AVs. The presented scenarios are often variations of the trolley problem asking the participant as an outside observer to choose between undesirable option such as killing two passengers or five pedestrians. The data indicates some global tendencies: people support minimizing loss of life and protecting children, favoring the fit and wealthy, and sacrificing people who are old, overweight, or homeless. The results also showed broad differences in relative preferences when comparing participants in different countries (e.g., the preference for sparing younger people rather than older ones is much higher for countries in the Southern cluster compared to the Eastern cluster). The implication is that "developing global, socially acceptable principles for machine learning" (Awad et al., 2018, p. 59) should be approached with great caution (Kochupillai, Lütge & Poszler, forthcoming). The findings from the study indicates that it is more effective to draw attention to the prohibition of discriminatory decision-making.

Past literature has extensively debated dilemma situations (e.g., unforeseen and unavoidable accidents) with reference to the famous trolley cases. The ideal is to avoid such situations in which accidents are unavoidable in the first place; for example, the lateral position of AVs on a lane can be adjusted to tune the risk posed to all other traffic participants (e.g., how much room should be given to a bicyclist?). Therefore, we argue to move away from the debate around dilemma situations. Instead, a r**esponsible balancing of risk or estimated harm should be permitted** for AVs at all times. This balancing decision should not be based on personal features of individuals such as age or gender (Lin, 2016; Lütge, 2017). Instead, as the severity of injury increases in proportion to the kinetic energy, estimated harm could be quantified and balanced by more objective factors such as the type or speed of particular traffic participants and the impact angle under which a collision would occur (Geißlinger et al., 2020). Taking into account the type of road users would grant vulnerable traffic participants (e.g., pedestrians or cyclists) the same level of protection as other road users (European Commission, 2020a). Overall, the consideration of these factors could help achieve a "[g]eneral programming to reduce the number of personal injuries" (Lütge, 2017, p. 552).

Besides the unbiased vehicle's internal functioning, AVs should be human-centric (European Commission, 2020a). In particular, AVs should be **equally usable for and accessible to all individuals**, which requires a **non-discriminatory design**. For example, age or the presence of a disability is not always considered by automotive companies, leading to potential issues of discrimination. Therefore, levels of differing abilities need to be acknowledged (e.g., a young individual may have quicker reflexes

for executing requests than elderly people) and the systems need to be adapted accordingly for different users, so that everyone can benefit from this new technology (for more information see also Guideline 1).

Industry recommendations:

- Companies should test their vehicle's AI systems for unfair performance differences across skin tone, gender, age and other characteristics. The IEEE P7003 standard can serve as a baseline to address and eliminate issues of bias in the creation of algorithms.
- When a technology is launched into the market, companies should localize it using data and train the model using multiple diverse data sets that are location specific.
- The AI developing team should be as inclusive as possible to include the broadest group possible in terms of demographics such as ethnicity.
- A responsible balancing of risks and potential harm to reduce the number of personal injuries should be permitted for AVs without discriminating against personal characteristics. Instead, factors underlying the balancing could include the type or speed of particular traffic participants and the impact angle under which a collision would occur.
- The personalization of AVs should be accessible by design and as inclusive as possible (e.g., disabilities included). Before an AV is released onto the streets, companies should demonstrate their plans and actions that ensure customizing-options to their vehicles (e.g., possibility to take away seats or include a ramp for entering the vehicle with a wheelchair).

Policy recommendations:

- Consideration should be given to having ethics standards boards test and assess that the systems for AVs are working properly, fairly and in an unbiased manner.
- Consideration should be given to requiring carmakers to explain the procedures they have put in place to make their designs accessible and avoid biases before granting them authorization to sell their vehicles to the public.

6. Societal and environmental wellbeing – including sustainability and environmental friendliness and social impact

In terms of societal and environmental wellbeing, the Sustainable Development Goals adopted by all United Nations Member States can serve as a reference point. Goal 3 (to "[e]nsure healthy lives and promote well-being for all at all ages") and goal 11 (aiming to "[m]ake cities and human settlements inclusive, safe, resilient and

sustainable") are particularly relevant to this topic (United Nations, 2015, p. 14). Companies and policymakers in the automotive sector should focus on meeting the following objectives:

1. **Increased public health and mobility**: AVs can improve society's health by avoiding fatalities that are due to human error (Bartneck et al., 2019). This is in line with Vision Zero, which states that eventually no one will and shall be killed or seriously injured within the road transport system (Ministry of Transport and Communications, 1997). The introduction of AVs could offer greater mobility solutions for a major part of society that is mobility-impaired, whether the elderly, young (without a driving license) or those who were otherwise unable to drive (BCG, 2017; WEF, 2018). This could positively affect mental health (e.g., due to feeling less dependent on others) and create a more inclusive society (Lim & Taeihagh, 2018). These benefits, however, can only be realized if safety and diversity standards are adhered to (for more information see Guideline 2 and 5).
2. **Better traffic flow**: AVs could reduce congestion and delays (e.g., during peak hours) and improve traffic flows and efficiency, especially when combined with shared mobility options. For example, using a traffic simulation model for Boston, it was found that the simulations that had included AV technology yielded less congestion, shorter travel times and more street space and (BCG, 2017; WEF, 2018). These benefits stem mostly from AVs' connectivity to external communication networks so that data can be managed and distributed in real time enabling methods such as platooning (Lim & Taeihagh, 2018). However, if not managed properly, it could also increase traffic flow and generate inefficiencies of uncoordinated traffic (Joaquin Acosta, 2018a). Proactive measures such as adopting a fitting physical and digital infrastructure, could improve the existing traffic situation by at least 15-20% (Inframix, 2020).
3. **Decreased carbon emissions**: Widespread adoption of AVs could reduce environmental degradation through reduced emissions and energy consumption (BCG, 2017). This is especially true if unnecessary acceleration and braking is reduced (Lim & Taeihagh, 2018). A concrete action point for companies would be to design AI systems that reduce vehicles' CO2 emissions. For example, companies could offer by default an eco-driving mode with a speed average that minimizes emissions and avoids unnecessary acceleration or braking. Many of the benefits relating to the reduction of carbon emission can be achieved by combining AVs with other disruptive technologies such as the electrification of vehicles (BCG, 2020). In addition, promoting AV shared mobility could "lessen the environmental impact of passenger vehicles by decreasing the number of vehicles on the road" (Joaquin Acosta, 2018a, p. 3). A concrete action point for policymakers would be to facilitate research and development for solutions to combine AVs with other disruptive technologies (e.g., electrification, shared mobility).

While these potential benefits are substantial, there is also significant uncertainty about the net impact of introducing AVs. Many measures of benefits focus on improvements per vehicle-mile traveled (VMT). However, the increased mobility and convenience benefits will potentially lead to significant increases in VMTs, potentially leading to increased total pollution, congestion, and so forth, despite the per-VMT gains (Geary & Danks, 2019). Thus, as technology continually develops, companies and policymakers in the automotive sector should **follow a stepwise implementation process to ensure that introduction of AVs provides net benefits.** Moreover, this implementation process must be combined with a **simultaneous adaption of infrastructure (physical and digital).** "Needed structural improvements include dedicated lanes to separate AVs from other traffic, and sensors to enable self-driving cars to communicate with their environment" (BCG, 2020). Otherwise, if AVs enter traffic in an uncoordinated way and without a fitting infrastructure, traffic flow and other benefits may be degraded. Several projects of the EU Horizon 2020 program have been focusing on this challenge (e.g., CoEXist or Inframix) (European Commission, 2019).

Inframix: The Inframix project aims at developing a road infrastructure for mixed vehicle traffic flows. Therefore, physical and digital elements of the road infrastructure need to be designed, upgraded and adapted to prepare for the stepwise introduction of automated vehicles without jeopardizing safety, quality of service and efficiency. This includes the design of novel visual signs and electronic signals that inform about the road operator's control commands and are readable and understandable by both automated and conventional vehicles. Further objectives of the project are to develop hybrid-testing systems by merging infrastructure elements and vehicles on real roads with a virtual traffic environment as well as to create a Road Infrastructure Classification Scheme that assess the level of 'automation-appropriateness' (Inframix, 2020).

City planners, road operators and local authorities should use the findings of such projects to make informed decisions on where to roll out new mobility models and how to update their road network accordingly. Collaboration with multiple private-sector leaders and national agencies is key to fostering innovation and progress: "the success of AMoD [autonomous mobility on demand] will depend to a large extent on establishing close partnerships among mobility providers, infrastructure companies, and city authorities" (BCG, 2020).

Industry recommendations:

- When developing their products, automotive companies should consider integrating and providing benefits of increased public health and mobility, better traffic flow and decreased carbon emission.
- Manufacturers should offer by default an eco-driving mode with a speed average that avoids unnecessary acceleration or braking and thus reduces carbon emissions.
- When developing AVs, car manufacturers should try to integrate other disruptive technologies such as electrification and shared mobility.

Policy recommendations:

- Policymakers should follow a stepwise implementation process and concentrate on mixed traffic scenarios. Policymakers should promote the integration of AVs in existing transport systems instead of competition between them, for example, by prioritizing research and development of AV solutions for public and shared mobility.
- A simultaneous adaption of physical and digital infrastructure is essential (e.g., lanes that separate AVs from other traffic).
- In doing so, collaboration with multiple actors such as private-sector leaders and national agencies is key to fostering innovation and progress (e.g., make use of projects investigating differing mobility models).

7. Accountability – including auditability, measures of transparency, reporting of negative impact, and redress

The attribution of liability and responsibilities for AVs is a challenging issue. "The first step towards the creation of a culture of responsibility is the study, deliberation and **agreement on the different responsibilities of different stakeholders**" (European Commission, 2020a, p. 56). In case of accidents, the AV itself cannot be held morally accountable since it is not considered a moral agent (Gogoll & Müller, 2017). Responsible parties will instead be manufacturers, component suppliers, technology companies, infrastructure providers or car holders / drivers. Therefore, policymakers should clarify the concept of a producer as well as **review regulations on product liability** (e.g., European Commission, 2018). This will, of course, vary depending on the motor vehicle laws in place in different countries. When adapting existing regulations to AVs, regulators may have to choose between different liability regimes for different situations and levels of automation. For example, strict liability

concepts may mean that for AVs, the manufacturer can be held liable if the automated mode is switched on, whereas, if not, the driver is considered liable. On the other hand, one could argue that liability should move gradually from one actor to the next (e.g., from the car manufacturer to the driver) depending on the driver's level of autonomy and solo action. For guidance, regulators could look to **Law Labs** (Joaquin Acosta, 2018b). Law Labs are a proposed concept to experiment with different regulatory approaches for a given innovation (e.g., AVs), similar to how regulatory sandboxes experiment with innovations in controlled environments (operating under temporary regulatory exemptions). For example, traffic rules could be revised and it could be investigated under which circumstances AVs are allowed to not to comply with a traffic rule (European Commission, 2020a).

In order to provide clarity about the causes of accidents, companies in the automotive sector may want to consider the following issues:

Regularly conduct internal and external audits. In terms of internal audits, manufacturers should execute *continuous optimization and tests*. This is in line with the guidelines of the Ethics Commission on Automated and Connected Driving, which state that "[l]iability for damage caused by activated automated driving systems is governed by the same principles as in other product liability [...:] manufacturers or operators are obliged to continuously optimize their systems and also to observe systems they have already delivered" (Lütge, 2017, p. 553). In doing so, companies should conduct a *risk assessment* by listing factors that may increase risk and uncertainty regarding a vehicle's operation and by proactively implementing appropriate countermeasures. Risks may stem from the vehicle's technology (e.g., technical failure to transmit sensor data), the actions of other traffic participants (e.g., disobeying traffic law such as jaywalking), external circumstances (e.g., the state of the road, weather conditions) or the vehicle's driving behavior (e.g., speed).

Uber Crash with jaywalker: In 2018, a self-driving vehicle owned by Uber Technologies Inc. struck and killed a pedestrian who was walking her bicycle across a road at night in Arizona. The underlying reasons for the accident included software flaws, such as the inability to recognize jaywalkers, which contributed to the failure to calculate that the vehicle could potentially collide with the pedestrian until only 1.2 seconds before impact, at which point it was too late to brake (National Transportation Safety Board, 2019). These types of crashes highlight the importance of prior risk assessment (such as the potential of other participants to disobey traffic rules) and subsequent redress mechanisms.

In addition to adhering to internal standards and audit requirements, external test centers could perform conformity assessments and grant *certifications since* "independent assessment will increase trust and ensure objectivity" (European Commission, 2020b, p. 25). Common audit areas for a certification system are – similar to the AI4People requirements – autonomy and control, fairness, transparency, reliability, security and data protection. Expected benefits of an AI certification would be trust in the application, orientation for customers and developers, fulfillment of norms such as cybersecurity and data security and comparable market equality (IAIS, 2019).

Implement explicit measures of transparency. These transparency measures should pertain to the development as well as the operation of AVs. Before operation, during the development phase, companies should **retain records and data** including data sets used for training (e.g., selection process) and document the programming and training methodologies. This is particularly important if authorities seek to review the underlying logic of a system and inspect relevant documentation (European Commission, 2020b). Also, during operations, relevant information should be recorded through **logging mechanisms and black boxes** integrated into AVs (Lütge, 2017). Regarding the black box, companies could consider an event data recorder and data storage system for AVs that record data of "the system status, occurrence of malfunctions, degradations or failures in a way that can be used to establish the cause of any crash and to identify the status of the automated/autonomous driving system and the status of the driver" (UNECE, 2019, pp. 3-4). These measures will ensure that the functioning and actions of AVs are explainable in retrospect. Overall, "[i]nternational standardization of the […] documentation (logging) is to be sought in order to ensure the compatibility of the logging or documentation obligations as automotive and digital technologies increasingly cross national borders" (Lütge, 2017, p. 555). For example, the upcoming regulation on automated lane keeping systems will determine what events are recorded by data storage systems for AVs (e.g., emergency maneuvers, failures) (UNECE, 2020b). IEEE P7001 provides such a standard for the transparency and accountability of autonomous systems so that the reasons why a technology makes certain decisions can be determined (IEEE, 2019). Similarly, SAE J3197 aims to govern data element definition, to provide a minimum data element set and common data output formats for an automated driving system data logger (SAE International, 2020).

Finally, **external communication and reporting of performance and negative impact** should be regularly required for companies in the automotive sector. Manufacturers and regulators should anticipate that individuals will want explanations when an AV's system did not perform as expected and intended. This is similarly stated in the ethical guidelines for trustworthy AI: "Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process" (High-level Expert Group on Artificial Intelligence,

2019, p. 18). In California, for example, Transportation Network Companies (TNCs) such as Lyft have to provide the California Public Utilities Commission with reports regarding zero tolerance complaints, violations and collisions of their vehicles on an annual basis (California Public Utilities Commission, 2020). Further information to disclose may be the tradeoffs made within algorithms, the number of past accidents put into context (e.g., relative number of accidents during test drives compared to total number of test drives) as well as safety measures initiated to counteract these accidents. Thereby data, algorithmic and AI literacy is improved (European Commission, 2020a).

Industry recommendations:

- Manufacturers should continuously conduct internal audits (e.g., assessing potential risks to the safe operation of AVs) and subsequently optimize their systems.
- Manufacturers should be transparent about the scope and process of their internal audits and risk assessments (e.g., space of conditions that are checked for).
- The internal audits should be complemented with regular external audits by independent test centers.
- Manufacturers should develop specific measures of transparency. This includes storing records and data of the underlying system logic (e.g., used training data sets) as well as logging mechanisms and black boxes (e.g., an event data recorder and data storage system) that document the actions of / in AVs during operation. The upcoming UN Regulation on Automated Lane Keeping Systems can serve as a baseline for vehicle manufacturers to develop appropriate data storage systems for AVs. SAE J3197 and the IEEE P7001 standard can serve as a baseline to address requirements for transparency and accountability of autonomous systems.
- Companies should transparently communicate and report performance and negative impacts of AVs (e.g., number of collisions, tradeoffs within algorithms).

Policy recommendations:

- Regulators should adapt laws and regulations concerning AVs and liability as the technology continues to develop. Regulators should clarify where responsibility lies in certain situations and ensure that privacy and cybersecurity damages are taken into account.
- Policymakers should consider establishing test centers that regularly request that companies perform conformity assessments and provide certifications.

CONCLUSION

This paper provides practical recommendations for the automotive sector to deal with ethical issues regarding: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing as well as accountability. By doing so, this paper distinguishes between policy and industry recommendations in suggesting first steps to be taken by both policymakers and companies. The following list summarizes all recommendations. In the future, we encourage stakeholders in the automotive sector to rely on these recommendations to determine relevant actions and to ensure that AVs comply with ethical principles.

AI4PEOPLE PRACTICAL RECOMMENDATIONS FOR THE AUTOMOTIVE SECTOR

Underlying Fundamental Rights

Policy recommendations:

- Relevant fundamental rights to be considered in the field of autonomous driving are: human dignity, right to self-determination and liberty, right to life and security, right to protection of personal data, right to equality and non-discrimination as well as the right to explanation.
- It must be realized that there will be no technologies or policies that maximize all fundamental rights for everybody simultaneously. There will always be trade-offs. Therefore, policymakers and legislators should decide which fundamental rights are to be prioritized in particular situations.
- In doing so, policymakers and legislators should cooperate with multiple stakeholders to obtain necessary information for executing an evaluation and subsequent agreement on compromises and prioritization.

1. Human agency and oversight

Industry recommendations:

- There should be a conditional override option allowing the control to be handed back to the driver. The admissibility of an override function depends on the level of automation of the AV (up to level 3: at any time; level 4: corresponding to safety mechanisms of an AV; level 5: not required) as well as on the state and behavior of the driver (e.g., impaired ability).
- AVs should continuously assess and monitor the driver's attentiveness and ability to intervene. Before operation, the AV could pose control questions to the driver (e.g., did you ingest any drugs or alcohol?); during operation, the AV could use sensors and biometric technology to do so. The upcoming UN Regulation on Automated Lane Keeping Systems can serve as a baseline for car manufacturers to develop appropriate driver attentiveness recognition systems.
- Handover should correspond to the driver's capabilities. Therefore, AVs could learn about drivers' capabilities and response times during operation from aggregated data and adjust the vehicle's parameters accordingly (respecting a safe minimum time response).
- Companies should provide documentation that justifies their particular handover window.
- Training programs should be tailored to different demographic groups and exhibit minimum elements that should be regarded in a training curriculum (e.g., limitations and capabilities of AVs) based on findings of recent studies.
- AVs should offer a 'training mode', for the first kilometers to train drivers on the AV's functioning.
- External human-machine interfaces should clearly communicate about the vehicle's motion intention and awareness of other traffic participants to humans outside the vehicle.

Policy recommendations:

- Policymakers should finalize what constitutes acceptable and legitimate override functions and define applicable situations for activation.
- Policymakers should determine standards for drivers' monitoring, training requirements, handover routines and external human-machine interfaces. These standards should be as global as possible.
- Policymakers in each jurisdiction should consider designating an organization in each jurisdiction that is allowed to look at the code and data within the AV in the event of a fatal accident involving an AV or a corresponding legal proceeding.

2. Technical robustness and safety

Industry recommendations:

- The prime requirement of AVs should be safety.
- In addition to threats that relate to conventional vehicles, manufacturers of AVs should particularly focus on cybersecurity threats. In doing so, companies need to comply with regulations for cybersecurity management systems. SAE J3061 could serve as a guideline to design cybersecurity into AVs throughout the entire development life cycle process.
- In terms of general functionality and safety, vehicles need to pass an official test that assures the system's accuracy, reliability and adequacy of its fallback options. The SAE Driving Safety Performance Assessment Metrics and the IEEE P7009 standard could serve as a baseline to design fail-safe mechanisms of autonomous and semi-autonomous systems.

Policy recommendations:

- Regulations need to be developed that reflect consensus on the method by which to grant approval to a vehicle's cybersecurity management system.
- Policymakers need to work with industry experts to develop a standardized test for the general functionality and safety of AVs to assure the system's accuracy, reliability and adequacy of its fallback options. This test could serve as a basis for the approval of AVs for sale to consumers.
- Policymakers need to collaborate with industry experts to determine and harmonize the definition of a 'safe condition' / 'minimal risk condition', the corresponding 'minimum risk maneuvers', and the circumstances in which such maneuvers should be executed. In doing so, SAE J3016 could serve as a baseline.

3. Privacy and data governance

Industry recommendations:

- Manufacturers should follow a strict privacy and data governance policy that includes transparency and communication to users, requesting affirmative consent and allowing limited sharing of data with third parties (including governments). In doing so, companies should comply with applicable standards and regulations such as the GDPR, the ePrivacy directive or the IEEE P7002.
- Before transmitting personal information from an AV to third parties, steps must be taken to ensure that it cannot be traced back to an individual.
- • Manufacturers should implement data protocols defining who can have access to data under which conditions.

Policy recommendations:

- Before receiving AV data, policymakers need to make clear what types of AV data they are seeking and how that data will enable them to improve public safety or some other legitimate public purpose (e.g., improve infrastructure, traffic flow and law compliance).
- At the EU level, building on Article 3(c) of Directive 2010/40/EU, consideration should be given to expanding the list of events and relevant traffic information that should be communicated free of charge.

5. Diversity, non-discrimination and fairness

Industry recommendations:

- Companies should test their vehicle's AI systems for unfair performance differences across skin tone, gender, age and other characteristics. The IEEE P7003 standard can serve as a baseline to address and eliminate issues of bias in the creation of algorithms.
- When a technology is launched into the market, companies should localize it using data and train the model using multiple diverse data sets that are location specific.
- The AI developing team should be as inclusive as possible to include the broadest group possible in terms of demographics such as ethnicity.
- A responsible balancing of risks and potential harm to reduce the number of personal injuries should be permitted for AVs without discriminating against personal characteristics. Instead, factors underlying the balancing could include the type or speed of particular traffic participants and the impact angle under which a collision would occur.
- The personalization of AVs should be accessible by design and as inclusive as possible (e.g., disabilities included). Before an AV is released onto the streets, companies should demonstrate their plans and actions that ensure customizing-options to their vehicles (e.g., possibility to take away seats or include a ramp for entering the vehicle with a wheelchair).

Policy recommendations:

- Consideration should be given to having ethics standards boards test and assess that the systems for AVs are working properly, fairly and in an unbiased manner.
- Consideration should be given to requiring carmakers to explain the procedures they have put in place to make their designs accessible and avoid biases before granting them authorization to sell their vehicles to the public.

6. Societal and environmental wellbeing

Industry recommendations:

- When developing their products, automotive companies should consider integrating and providing benefits of increased public health and mobility, better traffic flow and decreased carbon emission.
- Manufacturers should offer by default an eco-driving mode with a speed average that avoids unnecessary acceleration or braking and thus reduces carbon emissions.
- When developing AVs, car manufacturers should try to integrate other disruptive technologies such as electrification and shared mobility.

Policy recommendations:

- Policymakers should follow a stepwise implementation process and concentrate on mixed traffic scenarios. Policymakers should promote the integration of AVs in existing transport systems instead of competition between them, for example, by prioritizing research and development of AV solutions for public and shared mobility.
- A simultaneous adaption of physical and digital infrastructure is essential (e.g., lanes that separate AVs from other traffic).
- In doing so, collaboration with multiple actors such as private-sector leaders and national agencies is key to fostering innovation and progress (e.g., make use of projects investigating differing mobility models).

7. Accountability

Industry recommendations:

- Manufacturers should continuously conduct internal audits (e.g., assessing potential risks to the safe operation of AVs) and subsequently optimize their systems.
- Manufacturers should be transparent about the scope and process of their internal audits and risk assessments (e.g., space of conditions that are checked for).
- The internal audits should be complemented with regular external audits by independent test centers.
- Manufacturers should develop specific measures of transparency. This includes storing records and data of the underlying system logic (e.g., used training data sets) as well as logging mechanisms and black boxes (e.g., an event data recorder and data storage system) that document the actions of / in AVs during operation. The upcoming UN Regulation on Automated Lane Keeping Systems can serve as

a baseline for vehicle manufacturers to develop appropriate data storage systems for AVs. SAE J3197 and the IEEE P7001 standard can serve as a baseline to address requirements for transparency and accountability of autonomous systems.

- Companies should transparently communicate and report performance and negative impacts of AVs (e.g., number of collisions, tradeoffs within algorithms).

Policy recommendations:

- Regulators should adapt laws and regulations concerning AVs and liability as the technology continues to develop. Regulators should clarify where responsibility lies in certain situations and ensure that privacy and cybersecurity damages are taken into account.
- Policymakers should consider establishing test centers that regularly request that companies perform conformity assessments and provide certifications.

References

Archie, M., Gershon, S., Katcoff, A., & Zeng, A. (2018). De-anonymization of Netflix reviews using Amazon reviews. Retrieved from https://courses.csail.mit.edu/6.857/2018/project/Archie-Gershon-Katchoff-Zeng-Netflix.pdf

Autonomous Drivers Alliance (ADA) (2020). ADA Turing test. Retrieved from: https://ada.ngo/ada-turing-test

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. Nature, 563(7729), 59-64.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2019). An introduction to ethics in robotics and AI. Cham, Switzerland: Springer.

Boston Consulting Group (BCG) (2017). Making autonomous vehicles a reality: Lessons from Boston and beyond. Retrieved from https://www.bcg.com/de-de/publications/2017/automotive-making-autonomous-vehicles-a-reality.aspx

Boston Consulting Group (BCG) (2020). Can self-driving cars stop the urban mobility meltdown?. Retrieved from https://www.bcg.com/de-de/publications/2020/how-autonomous-vehicles-can-benefit-urban-mobility

California Public Utilities Commission (2020). Required reports TNCs must provide the CPUC. Retrieved from https://www.cpuc.ca.gov/General.aspx?id=3989

ESTECO (2020). Driving change for autonomous vehicles. Retrieved from https://mcusercontent.com/e18919a10879a5f50c06081a5/files/fb6d6b93-86c0-4bec-9bb0-0161a0629e09/WhitePaper_ADAS.pdf?utm_source=mailchimp&utm_campaign=0300efc2e1f0&utm_medium=page

Eurocybcar (2019). Cybersecurity test for cars. Retrieved from https://eurocybcar.com/en/

Europawire (2019). TomTom part of EU Data Task Force's proof of concept to make roads in EU safer. Retrieved from https://news.europawire.eu/tomtom-part-of-eu-data-task-forces-proof-of-concept-to-make-roads-in-eu-safer-20943847/eu-press-release/2019/06/05/11/18/39/73674/

European Commission (2013). REGULATION (EU) No 886/2013. Official Journal of the European Union. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013R0886&from=EN

European Commission (2018). Liability for emerging digital technologies. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018SC0137&from=en

European Commission (2019). Automated road transport – On the way to connected and automated mobility. Retrieved from https://ec.europa.eu/inea/sites/inea/files/art_brochure-2019.pdf

European Commission (2020a). Ethics of connected and automated vehicles. Retrieved from https://ec.europa.eu/info/sites/info/files/research_and_innovation/ethics_of_connected_and_automated_vehicles_report.pdf

European Commission (2020b). On artificial intelligence – A European approach to excellence and trust. Retrieved from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Cybersecurity Certification Group (ECCG) (2020). The EU cybersecurity certification framework. Retrieved from https://ec.europa.eu/digital-single-market/en/eu-cybersecurity-certification-framework

European Data Protection Board (EDPB) (2020). Guidelines 1/2020 on processing personal data in the context of connected vehicles and mobility related applications. Retrieved from https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_202001_connectedvehicles.pdf

European Union Agency for Cybersecurity (ENISA) (2019). ENISA good practices for security of mart cars. Retrieved from https://www.enisa.europa.eu/publications/smart-cars/at_download/fullReport

Federal Ministry of Transport and Digital Infrastructure (BMVI) (2015). Minimum risk manoeuvres. Retrieved from https://wiki.unece.org/download/attachments/27459841/ACSF-04-07%20%20%28D%29%20-%20ACSF-MRM.pdf?api=v2

Federal Ministry of Transport and Digital Infrastructure (BMVI) (2017). Ethics Commission automated and connected driving. Retrieved from https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile

Florentine, E., Ang, M. A., Pendleton, S. D., Andersen, H., & Ang Jr, M. H. (2016). Pedestrian notification methods in autonomous vehicles for multi-class mobility-on-demand service. In Proceedings of the Fourth International Conference on Human Agent Interaction (pp. 387-392). New York, NY: Association for Computing Machinery.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Lütge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. & Vayena, E. (2018). AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707.

Forbes (2020). Overcoming racial bias in AI systems and startlingly even in AI self-driving cars. Retrieved from https://www.forbes.com/sites/lanceeliot/2020/01/04/overcoming-racial-bias-in-ai-systems-and-startlingly-even-in-ai-self-driving-cars/#2b1cc433723b

Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) (2019). Trustworthy use of artificial intelligence: Priorities from a philosophical, ethical, legal, and technological viewpoint as a basis for certification of artificial intelligence. Retrieved from https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_Thrustworthy_AI.pdf

Future of Privacy Forum (2017). Data and the connected car. Retrieved from https://fpf.org/wp-content/uploads/2017/06/2017_0627-FPF-Connected-Car-Infographic-Version-1.0.pdf

Geary, T., & Danks, D. (2019). Balancing the benefits of autonomous vehicles. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 181-186). New York, NY: Association for Computing Machinery.

Geißlinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2020). Autonomous driving ethics: From Trolley problem to ethics of risk. Working paper.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. Science and engineering ethics, 23(3), 681-700.

High-level Expert Group on Artificial Intelligence (2019). Ethics guidelines for Trustworthy AI. Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

IEEE (2019). Ethically aligned design – A vision for prioritizing human well-being with autonomous and intelligent systems. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf

Inframix (2020). Expected impact – a step by step introduction of automation. Retrieved from https://www.inframix.eu/expected-impact/

Joaquin Acosta, A. (2019). IoT international regulator challenges: The European approach. In C. H. Cwik, C. A. Suarez, & L. L. Thomson (Eds.), The internet of things (Iot): Legal issues, policy, and practical strategies (pp. 191-215). Chicago, IL: American Bar Association.

Joaquin Acosta, A. (2018a). Autonomous vehicles: A smart move? 24 essentials of a SWOT analysis policymakers need to consider. Berkman Klein Center for Internet and Society at Harvard University. Retrieved from https://cyber.harvard.edu/sites/default/files/2018-07/2018-07_AVs02_0.pdf

Joaquin Acosta, A. (2018b). Autonomous vehicles: 3 practical tools to help regulators develop better laws and policies. Berkman Klein Center for Internet and Society at Harvard University. Retrieved from https://cyber.harvard.edu/sites/default/files/2018-07/2018-07_AVs04_1.pdf

Kochupillai, M., Lütge, C., & Poszler, F. (forthcoming). Programming away human rights and responsibilities? The Moral Machine Experiment and the need for a more 'Humane' AVs Future. NanoEthics.

Kriebitz, A., & Lütge, C. (2020). Artificial intelligence and human rights: A business ethical assessment. Business and Human Rights Journal, 5(1), 84-104.

Leonhardt, T. (2018). Minimal risk maneuver. Retrieved from https://www.ko-haf.de/fileadmin/user_upload/media/abschlusspraesentation/12_Ko-HAF_Minimal-Risk-Maneuver.pdf

Lim, H. S. M., & Taeihagh, A. (2018). Autonomous vehicles for smart and sustainable cities: An in-depth exploration of privacy and cybersecurity implications. Energies, 11(5), 1062.

Lin, P. (2016). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.) Autonomous driving (pp. 69-85). Berlin, Heidelberg, Germany: Springer.

Lütge, C. (2017). The German ethics code for automated and connected driving. Philosophy & Technology, 30(4), 547-558.

Manser, M. P., Noble, A. M., Machiani, S. G., Shortz, A., Klauer, S. G., Higgins, L., & Ahmadi, A. (2019). Driver training research and guidelines for automated vehicle technology. Retrieved from https://vtechworks.lib.vt.edu/bitstream/handle/10919/95178/01-004_Final%20Research%20Report_Final.pdf?sequence=1

Merat, N., Jamson, A. H., Lai, F. C., Daly, M., & Carsten, O. M. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. Transportation research part F: traffic psychology and behaviour, 27(Part B), 274-282.

Ministry of Transport and Communications (1997). En route to a society with safe road traffic. Retrieved from https://trid.trb.org/View/512093

National Transportation Safety Board (2019). Vehicle automation report, Tempe, AZ, HWY18MH010. Retrieved from https://www.documentcloud.org/documents/6540547-629713.html

National Transportation Safety Board (2020). Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator. Retrieved from https://www.ntsb.gov/investigations/accidentreports/pages/har2001.aspx

NHTSA (2017). Automated driving systems 2.0. Retrieved from https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf

Pagallo, U., Aurucci, P., Casanovas, P., Chatila, R., Chazerand, P., Dignum, V., Lütge, C., Madelin, R., Schafer, B., & Valcke, P. (2019). AI4People on good AI governance: 14 priority actions, a SMART model of governance, and a regulatory toolbox. Retrieved from: https://ssrn.com/abstract=3486508

Research and Markets (2019). Analysis of driver monitoring systems, 2020 Edition. Retrieved from https://www.researchandmarkets.com/reports/4893877/analysis-of-driver-monitoring-systems-2020?utm_source=dynamic&utm_medium=GNOM&utm_code=c968tk&utm_campaign=1338270+-+2020+Analysis+of+the+Global+Driver+Monitoring+Systems+Market&utm_exec=cari18gnomd

Rouchitsas, A., & Alm, H. (2019). External human–machine interfaces for autonomous vehicle-to-pedestrian communication: A review of empirical work. Frontiers in psychology, 10(2757), 1-12.

SAE International (2018). J3016 – Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Retrieved from https://saemobilus.sae.org/content/J3016_201806

SAE International (2020). J3197 – Automated driving system data logger. Retrieved from https://www.sae.org/standards/content/j3197_202004/

SAE International – Vehicle Cybersecurity Systems Engineering Committee (2016). SAE J3061: Cybersecurity guidebook for cyber-physical vehicle systems. Retrieved from https://www.sae.org/standards/content/j3061_201601/

Secretary of State Washington (2020). Certification of enrollment – engrossed substitute senate bill 6280. Retrieved from http://lawfilesext.leg.wa.gov/biennium/2019-20/Pdf/Bills/Senate%20Passed%20Legislature/6280-S.PL.pdf?q=20200409103455

Techcrunch (2006). AOL proudly releases massive amounts of private data. Retrieved from https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

Tesla (2020). Using autopilot and full self-driving capability. Retrieved from https://www.tesla.com/support/autopilot

United Nations (1948). Universal Declaration of Human Rights. Retrieved from https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf

United Nations (2015). Transforming our world: The 2030 agenda for sustainable development. Retrieved from https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

United Nations Economic Commission for Europe (UNECE) (2019). Revised framework document on automated/autonomous vehicles. Retrieved from https://www.unece.org/fileadmin/DAM/trans/doc/2020/wp29/ECE-TRANS-WP29-2019-34-Rev2e.pdf

United Nations Economic Commission for Europe (UNECE) (2020a). Draft new UN regulation on uniform provisions concerning the approval of vehicles with regard to cyber security and of their cybersecurity management systems. Retrieved from https://www.unece.org/fileadmin/DAM/trans/doc/2020/wp29grva/GRVA-06-19r1e.pdf

United Nations Economic Commission for Europe (UNECE) (2020b). UN regulation on automated lane keeping systems is milestone for safe introduction of automated vehicles in traffic. Retrieved from https://www.unece.org/info/media/presscurrent-press-h/transport/2020/un-regulation-on-automated-lane-keeping-systems-is-milestone-for-safe-introduction-of-automated-vehicles-in-traffic/doc.html

Vox (2019). A new study finds a potential risk with self-driving cars: Failure to detect dark-skinned pedestrians. Retrieved from https://www.vox.com/future-perfect/2019/3/5/18251924/self-driving-car-racial-bias-study-autonomous-vehicle-dark-skin

Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. arXiv preprint arXiv:1902.11097.

Wishart, J., Como, S., Elli, M., Russo, B., Weast, J., Altekar, N., James, E., & Chen, Y. (2020). Driving safety performance assessment metrics for ADS-equipped vehicles. SAE International. Retrieved from https://www.researchgate.net/profile/Jeffrey_Wishart/publication/340652968_Driving_Safety_Performance_Assessment_Metrics_for_ADS-Equipped_Vehicles/links/5eb0a39e92851cb2677403ba/Driving-Safety-Performance-Assessment-Metrics-for-ADS-Equipped-Vehicles.pdf

World Economic Forum (WEF) (2018). Reshaping urban mobility with autonomous vehicles – Lessons from the City of Boston. Retrieved from http://www3.weforum.org/docs/WEF_Reshaping_Urban_Mobility_with_Autonomous_Vehicles_2018.pdf

2

BANKING & FINANCE

Authors

Nir Vulkan
Chairman Banking & Finance Committee, AI4People; Associate Professor of Business Economics at Saïd Business School, University of Oxford, UK

Aisha Naseer
AI Ethics Research Manager at Fujitsu Laboratories of Europe

Frank McGroarty
Chairman Insurance Committee, AI4People; Professor of Computational Finance and Investment Analytics; Director of Centre for Digital Finance at Southampton Business School, UK

Giulia Del Gamba
Digital and Innovation Policy Advisor at Intesa Sanpaolo

John Cooke
Chairman of the Liberalisation of Trade in Services Committee at TheCityUK

Lampros Stergioulas
Professor in Business Analytics at the University of Surrey, UK

Paul Jorion
Associate Professor of Ethics, Université Catholique de Lille, France

Raffaella Donini
Senior Manager, European Digital and Innovation policies at Intesa Sanpaolo

1. Executive Summary

AI technology has had an enormous impact on the Banking & Finance sector (B&F from hereon in). Moreover, this trend is only likely to continue. Already most credit check, KYC and AML decisions are now made by algorithms. Credit scores and therefore credit worthiness tests are also hugely faster and more accurate when carried out by algorithms trained on large data sets based on past decisions and outcomes (Khandani et al., 2010). Similarly, investing and trading continue to be disrupted by AI - in some markets more trade happens as results of orders put in by algorithms than by humans. The so-called "FinTech Revolution" which has disrupted and forced a rethink of existing paradigms within B&F, is strongly assisted by AI technologies (Noya, 2019); for example, many firms use AI to combat fraud (Stripe, 2020; Amazon Web Services, 2020; Chatfield, 2017; Thanendran, 2018). However, the mass adoption of AI over the past ten years has also brought up important concerns. When using these technologies, institutions now have to ask themselves fundamental questions – are the algorithms we are using fair and transparent? Are our customers adequately protected from risk? Is our users' data safe? Are these approaches sustainable? And crucially, is using AI worth it?

This committee believes that AI has the potential to effect a great amount of positive change within this important sector. More specifically, we believe AI can help address three of the most important challenges faced by finance nowadays: financial inclusion, financial literacy and financial wellbeing. Additionally, if AI is used responsibly, it can lead to higher revenue growth, cost efficiencies and a better customer experience. Conversely, poor use of AI may lead individuals to be less engaged with their finances, may propagate further discrimination and may foster exclusion. At the level of a firm, AI may be hugely cost inefficient or it may negatively impact their operational resilience if it is not deployed correctly, safely and efficiently.

We note that finance is already the most regulated area of industry. In many sectors of financial services, regulatory objectives are now set by global standard-setters (e.g. the Basel Committee), leaving precise implementation to supervisors in individual jurisdictions. This means that much regulation (whether prudential regulation or market conduct regulation) is already highly developed and potentially all-embracing. Thus, we believe that the development of wholly new regulation pertaining to the use of AI within financial services will be unlikely. Indeed, there is likely to be existing regulation, which in principle is technologically neutral. Instead, we believe the crucial question is how existing regulation (or the skills of regulators and supervisors) should be adapted to cater for the enhanced role of AI in previously existing modes of delivery of financial services. This ought to mean that further whole layers of regulation need not be added.

This could well prove fortunate, because adding too many layers of regulation would have anti-competitive effects, further increasing the barriers to entry for start-ups and resulting in unfair advantages accruing to big banks and tech giants who are now entering this space. Like many innovative areas in technology, AI will need to be fair, safe and user-friendly if it is to gain mass adoption. Part of this step will involve harmonised standards that, at a minimum, offer guidance and consumer protection and, ideally, provide a regulatory framework. When considering the structure of this framework, regulators may want to consider leveraging and adapting existing regulatory solutions, to help these new technologies to be used fairly and safely. However, new categories of risk may also emerge, in which case targeted regulatory remedies should be available in order to protect consumers, encourage healthy competition, and ensure market stability. Regulators often favour a principled, outcome-based approach for regulating fast paced and quickly innovating areas - with this, developed frameworks are adaptable and can evolve over time. Alternatively, given the huge range of variables and use cases for AI, it may equally be sensible to adapt a risk-based approach and consider the issues that arise on a case by case basis.

For this reason, in this document we consider the impact of AI technologies on the B&F sector in light of the seven key requirements for Trustworthy AI laid down in the Ethics Guidelines (High-Level Expert Group on Artificial Intelligence, 2019), the six key features provided by the European Commission White Paper on AI (European Commission, 2020), as well as the European Parliament Framework of ethical aspects of artificial intelligence, robotics and related technologies (European Parliament, 2020). Given existing regulation (for example, in Europe, the Markets in Financial Instruments Directive and Regulation, the General Data Protection Regulation, or the Capital Requirements Directive) we believe that the following five requirements should be addressed as priorities within the B&F sector:

1) **Fairness and non-discrimination**
2) **Technical robustness and safety**
3) **Transparency and explainability**
4) **Accountability**
5) **Human oversight**

These five principles are our main points of focus. The remaining principles may be important for other industries, or for AI in general, but we concentrate on the five above, as we believe they are the most applicable to our setting.

We begin by providing a general overview of AI and its role in the B&F sector. Then for each of the five principles we provide use cases, review the research and

literature produced by academia, banks and regulators, before supplying our recommendations. Finally, we summarise the recommendations both for firms and regulators.

A key issue throughout this document is that principles such as fairness, accountability etc do not have a universally agreed interpretation in the context of B&F. In other words, they are contestable. This document highlights some ideas and suggestions as to how to address this problem. A very useful comparison can be drawn between existing industrial ethics frameworks, and new AI ethical frameworks for that same industry. In particular, in the context of the ethics of AI in medicine, (Mittelstadt, 2019) argues that whilst emerging frameworks are seemingly aligned with traditional principles, language is being used that "hide[s] deep political and normative disagreement".

2. Overview of ai and its role in banking and finance

Artificial Intelligence is a broad term which captures a range of technologies. As such, it is somewhat difficult to pin an exact definition on it. In (Hofstadter, 1979), the author quips that "AI is whatever hasn't been done yet." For our purposes, we can understand AI to mean those technologies which emulate those human capabilities we value, but which we traditionally understand to be beyond the reach of computing devices. This includes activities such as understanding the "meaning" of pictures, videos and audio, rather than just treating them as digital signals, as well as problem solving and reasoning about unseen situations. Often, we like to think of AI as the study of intelligent agents – software we can delegate tasks to in order to achieve our goals.

AI has co-evolved with the inception of computing in the 1940s, and has been an active area of interest throughout the 20th and 21st centuries. Since 2011, however, there has been an explosion of both practical results, as well as popular interest, in the subject. This is largely due to one strand of AI becoming increasingly prevalent – Machine Learning (ML). Whilst the techniques of ML have existed since the 1980s, they only really came into their own when researchers showed how to efficiently implement them on Graphics Processing Units – relatively inexpensive, commodity hardware. Since then, AI has touched almost every industry and now affects us all on a day-to-day basis – B&F is no exception. B&F is a vast, complicated industry. In the UK, for example, the financial services industry is responsible for almost 7% of economic output (Rhodes, 2019).

Banking and finance differ from other industries covered by the other AI4People panels, in that the different parts of the sector, for example trading, are themselves trillion-dollar industries (Pound, 2019). The sector is already highly regulated, especially after

the 2007-2008 financial crisis. This increased regulation carries significant costs for the industry, with B&F firms spending considerable amount of their resources on compliance.

AI is already being used in many areas across B&F, including algorithmic trading, robo-advisors, fraud detection and automated loan decisions, and has the potential to effect further transformation in banking and finance. The B&F sector has unique characteristics which may make the use of AI difficult to regulate: in particular, the sector's need to focus on risk management should always be held in mind when reviewing new technologies. Furthermore, the industry is often held to higher societal standards than other industries, especially since the 2007-2008 financial crisis. It is important regulators get the balance right between regulation and innovation, whilst reassuring the public that the institutions taking care of their money are safe.

It is illustrative to study the growth of FinTech, having disrupted traditional financial services, and largely assisted by AI technology. Figure 1 below (Vulkan, 2019a) represents the functional framework for understanding finance and illustrates the relationship between the services provided by financial institutions and the systems and structures that form the foundation for these services.
In the UK a survey by the Bank of England and FCA on machine learning in financial

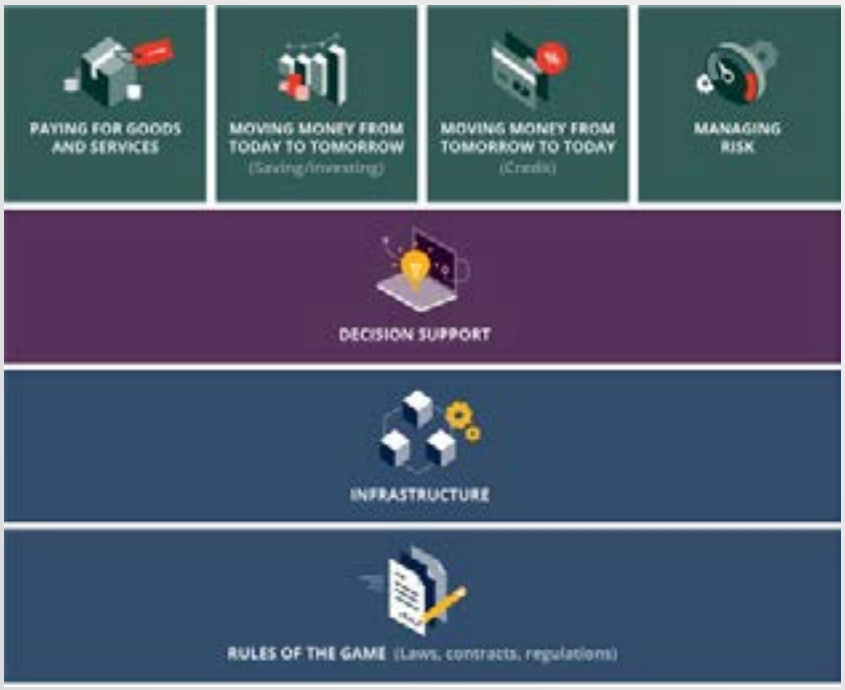


Figure 1: A functional view of the B&F sector

services found that two thirds of respondents already used machine learning in some form (Bank of England, and Financial Conduct Authority, 2019). Officials on Wall Street plan to use artificial intelligence systems and machine learning to monitor the stock markets and predict patterns of fraud (Reuters, 2016). The potential of AI should not be underestimated – according to some research the use of AI across the economy could boost the UK's labour productivity by 25% by 2035 and add £650bn to UK gross value added (GVA) (International Regulatory Strategy Group, and Accenture, 2019).

Figure 2 below (Pinsent Masons, and Innovate Finance, 2019) illustrates some areas where AI is already being used in the sector. Taking fraud detection as an example, it is clear that AI can be much more efficient than humans. ML technologies can be used to analyse large volumes of data and detect irregular patterns much faster than a person could. As such, AI makes it substantially easier for officials to examine large amounts of potentially suspicious data and patterns (OECD, 2017).

Figure 2: Common uses of AI in the financial services sector

In (International Regulatory Strategy Group, and Accenture, 2019), the authors list the key benefits of AI for industry participants and consumers as: higher revenue growth, increased cost savings, improved customer experience and better risk mitigation. However, AI is not without its risks. In the next section, we will explore the potential dangers of AI, discuss their mitigations and provide recommendations for both industry and policymakers on how AI can be Trustworthy within the B&F industry.

Before proceeding, we note that the B&F sector is not isolated from other sectors of industry – there are deep connections between B&F and the remaining six sectors outlined by AI4People. Thus, when talking about risks, mitigations and recommendations for B&F, it is important to be cognisant of how these elements link with corresponding ideas in other sectors. In particular, given the porous barriers within the collective ecosystem of sectors, when we make recommendations, we should consider their impact not just on B&F, but on industry as a whole.

3. Analysis and Recommendations

3.1. Fairness and non-discrimination

Fairness within artificial intelligence is an active area of research, and there are many proposals both for detecting/defining fairness (Sahil et al., 2018) and for mitigating discrimination bias (Mehrabi et al., 2019) in machine learning models.

Defining "fairness" in itself is difficult, since there is not a unique intuitive notion of what fairness in decision making should be. In the context of B&F, what is fair is still very much uncertain. Proposals have been put forward to measure "average" discrimination between groups of people (e.g. the different percentage of loans granted to women and to men), and also to provide individual notions of fairness (two similar individuals should be given the same decision). Both of the above notions of fairness (i.e. group fairness and individual fairness) can be further broken down into a plethora of different concepts, which are, in general, not mutually compatible (Kleinberg, 2016). Additionally, different mitigation techniques have been developed in order to build models that meet specific fairness requirements.

Thus, a lot of work has already been done in the last few years to understand discrimination in machine learning, measure it and mitigate its effects. However, caution must be taken in order to understand what notion of fairness is most suitable for the use case in question. This problem goes beyond the technical implementation of tools and procedures, and as such, calls for joint contributions from AI specialists, legal experts, firms, governments and even philosophers.

Fairness and non-discrimination are universally considered to be key requirements for AI systems. The aim of the non-discrimination principle is to allow all individuals an equal and fair prospect to access opportunities available in a society. Individuals who are in similar situations should receive similar treatment and should not be treated less favourably due to their protected characteristics. Indirect discrimination is present when certain characteristic or factor occurs more frequently in the population groups against whom it is unlawful to discriminate. Since algorithmic decision-making systems may be based on correlations, there is a risk to perpetuate or exacerbate indirect discrimination through stereotyping, when differential treatment cannot be justified (Council of Europe, Committee of experts on Internet MSI-NET, 2017). Financial data is prone to bias and imbalance (Zhang and Longsheng, 2019) and putting fairness principles into practice in industrial processes is an open issue. In (Saxena, 2019), the authors conducted online experiments to examine people's perceptions around fairness definitions and found that one specific definition, calibrated fairness, tended to be preferred over two other possibilities. However, these experiments were conducted solely around one problem (loan decisions) and cannot be taken as the final word on the matter. Additionally, (Green and Chen, 2019) explored the algorithm-in-the-loop paradigm, and demonstrated in an experiment how human predictions, aided by a risk assessment algorithm, exhibited a great deal of bias.

Bias

Shortly after Apple released its own credit card in 2019, a number of people reported that women were receiving lower credit limits than men with similar financial backgrounds (Knight, 2019). Goldman Sachs, the issuing bank, claimed that a third party had audited the credit algorithm for bias and stated that it did not depend on protected attributes such as race, age or sex, but only on their "creditworthiness". However, it is clear that a person's sex can be inferred from their past transactions, and as such, their algorithm may have indeed (accidentally) learned to discriminate against women.

Even if a team of developers have the best intentions, bias can easily slip into a model. This can happen in a multitude of ways - from how the data was collected and aggregated, to an oversight of a developer, to the algorithm used in the training of the model itself.

How can we eliminate such bias from our models, when we may not even be aware that we are introducing it in the first place? One appealing

approach is that of Counterfactual Fairness (Kusner et al., 2017). Suppose your algorithm has made a prediction, and you want to determine if that prediction is "fair" with regard to some protected attribute. Counterfactual Fairness asks you to imagine a parallel world, where that attribute has been changed and to see whether your algorithm makes a different prediction. Far from being an abstract thought experiment, (Kusner et al., 2017) actually introduced a debiasing algorithm to formally implement this notion, and it offers promising results in terms of largely retaining accuracy of the original algorithm, whilst minimising unfairness.

However, Counterfactual Fairness is only but one method for removing bias from the decision making process - for a survey on bias and techniques for its removal, (Mehrabi et al., 2019) is a good reference.
Despite its appeal, Counterfactual Fairness comes not without flaws. Among others, the fact that it requires very strong assumptions about the causal relationships among variables at play, some of which are not even falsifiable. For this reason, the idea itself of "counterfactuals" is still highly debated by the scientific community.

Trying to eliminate bias in algorithms is currently an art, rather than a science. To this end, there are a number of routes one can take. Whilst fairness is not a universally agreed upon concept, there are a number of metrics which have been developed in order to judge whether an algorithm is acting "fairly" or not. When data scientists are developing algorithms and training models, they should calculate these metrics, to provide some insight as to whether they are being unfairly discriminatory in any way.

Recommendations:

- Instead of trying to entirely eliminate bias, one should learn to manage it instead. This includes being able to identify the potential for bias in models and datasets, as well as understanding the types of algorithms that can mitigate its effect.
- Document and publish an ethical code of conduct policy to promote non-discrimination principles (CSSF, 2018);
- Include bias assessment and mitigation into the AI project pipeline. This can be done, e.g., by means of additional steps such as in (Castelnovo et al., 2020): a careful exploration and understanding of the problem at hand and of the available data in order to identify the possible unfair impact on the user; monitor different metrics of fairness along the train, validation and test phases; implement, when necessary, available strategies in order to mitigate the bias; evaluate results along multiple dimensions and compare the implemented strategies; re-think some of the steps or iterate the process, when necessary;

- Constantly monitor the performance of the model not only overall but also at the level of potentially discriminated groups;
- Since trade-offs between fairness and fidelity (or accuracy) still persist, which ML model should be used for the problem in question significantly depends on the context and its business domain.

3.2. Technical Robustness and Safety

As alluded to in the previous section, minimising loss and maximising accuracy is not the be-all and end-all of ML. There is a classic (probably apocryphal) example (Branwen, 2011) that illustrates this point nicely – allegedly an American government agency wanted to train a neural network to differentiate friendly tanks from enemy tanks. They followed the basic recipe for machine learning, taking care to separate their data into train and test sets, before running the training algorithm on the dataset. The model was a huge success – it could successfully differentiate between friendly and enemy tanks with close to 100% accuracy. However, when they decided to further test the model with new photos, it was useless – seemingly performing no better than random chance. It then emerged that the pictures of friendly tanks had been taken on a sunny day, whilst the pictures of enemy tanks had been taken on a cloudy day. The model had simply learned how to tell how bright a photo was.

Whilst the above story may only contain a grain of truth, it does encapsulate some important points. Primarily, ML isn't just an optimisation game – it is a game of balance. There is a vast difference between a model which performs well in development, and a model which performs well in production. There are two key ways in which we can try and ensure that an ML deployment is a successful one. First, a model should be statistically sound. That means evaluating it with respect to a range of metrics, rather than just picking and choosing the ones that make it look like it is performing well. In particular, the number of true/false positives/negatives of a model should be closely analysed. Additionally, before even training a model, the underlying data must be hand-analysed and understood, to ensure the engineers appreciate any "quirks" in the data. Note that this may require the use of domain-specific experts. Second, there are many regularisation techniques which can be used to improve model robustness. There are far too many individual techniques to list here, but resources such as (Goodfellow et al., 2016) offer a good coverage of the most important techniques.

Building on this point, as AI is software, we stress that the techniques and frameworks from traditional software testing should firmly remain one of the core steps in any AI deployment. This includes all the usual exercises of unit testing, integration testing and

acceptance testing – these activities provide assurance that the system not only works, but functions as a cohesive whole. However, in the context of ML, we can provide further assurance through the subfield of neural network verification. Whilst this field is relatively new, a number of different approaches have been suggested, many of which seem promising. Again, we believe it is up to regulators to track the progress within this field and determine whether any of these methods are appropriate and/or necessary for the subsector in question.

Finally, it is worth touching on two more topics - model data "leakage" and adversarial attack vectors.

It has been shown (Fredrikson et al., 2015; Song et al., 2017; Carlini et al., 2019) that ML models are capable of "leaking" their training data. That is, given a pretrained model, they show how examples of the training data can be reverse-engineered from it in certain circumstances. This has the damaging potential to leak private and confidential client information. As such, engineers should be aware of how to minimise this possibility. One emerging, promising field is that of differential privacy (Ji et al., 2014). This tries to ensure that the information from the model in question does not change significantly if one training example is added or removed. This has the knock-on effect of it being more difficult to extract the data of individuals from the model. As ML becomes more prevalent and firms contract third-parties to train models for them, ideas like this will become increasingly important. As such, it is crucial that both engineers and regulators track the development of ideas in this subfield closely, in an attempt to mitigate data leakage.

Adversarial data (Szegedy et al., 2014; Goodfellow et al., 2015) consists of examples which look "normal" to humans, but which neural networks consistently classify incorrectly. For instance, (Goodfellow et al., 2015) create a picture of a panda, which is consistently classified as a gibbon. Raising the stakes, (Sitawarin et al., 2018) show how road signs can be subtly modified to trick autonomous vehicles into following the wrong traffic rules, whilst (Sharif et al., 2016) show how to modify common eye-glasses to impersonate another individual from the viewpoint of a facial recognition algorithm. Perhaps even more surprisingly, (Hendrycks et al., 2019) provide natural (i.e. non-engineered) images which are repeatedly classified incorrectly. Thankfully, work has been done which aims to mitigate the effect of such attacks (Madry et al., 2017) and we believe engineers should be aware of and know how to implement such measures.

Recommendations:
Engineers should be encouraged to take on training in statistical analysis of models and the datasets – in particular statistical verification and the risks of overfitting;

- Traditional software testing should remain a critical part of any AI deployment and new ideas from neural network verification should also be integrated into the process;
- Engineers should be aware of model data "leakage", as well as adversarial attack vectors and receive training in the latest news in this area and learn how to mitigate such risks (as much as possible).

3.3. Transparency and explainability

Another crucial line of thought from (High-Level Expert Group on Artificial Intelligence, 2019) is the following - "*Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system*". Herein lies two important points – letting users understand the extent of their interaction with AI systems and providing engineers and regulators with the training to build, regulate, and understand these systems in an appropriate way.

We believe it is important for humans to know whether they are interacting with an algorithm or another person. If a client or customer is talking to a chatbot, say, then said chatbot should make clear at the start of the interaction that it is not human. Equally, if a decision has been taken with the help of an AI algorithm, this information should be disclosed during the process. This provides two core benefits – it preserves and reinforces the role of human autonomy in our interaction with AI systems, and also gives humans the capacity to challenge decisions made by AI. This helps emphasise that AI is simply a tool in any business process, rather than an unchecked, autonomous agent. We will touch on another side of this same coin, *explainability*, when we come to talk about fairness.

Explainability

Historic approaches to AI (pre-1990) largely relied on symbolic, logic-based techniques. A key example of this is that of "expert systems" - these encapsulated domain-specific knowledge and used deduction rules to form conclusions. For example, if you have the rule "If a patient has a high temperature and a cough, then there is an 80% chance they have the flu", along with the information that patient A has a high temperature, along with a cough, then we can conclude that patient A has an 80% chance of having the flu. Expert systems would have thousands of these rules which could be used in succession to reach advanced conclusions. Such systems offered immediate explainability – if a decision or prediction had been made and you wanted to understand where it came from, you can simply

trace back the rules used to make such a decision/prediction until you have reached the desired level of granularity of understanding. Whilst expert systems are not used widely nowadays (they are labour intensive to create, require the input from domain-specific experts, and were never hugely successful), their influence can be seen in modern knowledge-based systems, such as Google's Knowledge Graph, IBM's Watson, as well as in personal AI assistants such as Siri, Cortana and Alexa.

Conversely, modern Artificial Intelligence largely uses methods rooted in Machine Learning and statistics, rather than logic. In particular, one of the major successes of AI in the 21st century has been the explosion of deep learning methods. These techniques are alarmingly effective at capturing domain knowledge, by learning patterns in training data. However, this "knowledge" is encapsulated in many "layers" of "neurons", and their individual behaviour is defined by an array of numbers called the "parameters" of the neuron. State of the art models often have millions and billions of parameters in total. As such, it is almost impossible to ask questions like "what does the number -0.29 mean in neuron 47 in layer 6 of the model?" - it is difficult to understand the context of individual parameters. Put differently, whilst machine learning does effectively capture intuition and knowledge within a domain, it is hard to extract this knowledge in a human-understandable format – they are black-box models.

Two recently proposed but already popular model-agnostic explainability tools are LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). While their internal mechanisms are different, the general idea is the same: building surrogate models capturing the changes in the prediction given the changes in the input. They perform a sensitivity analysis by tweaking the input slightly and then testing the resulting changes in prediction. For example, if the model prediction does not change much by tweaking the value of a variable, that variable for that particular data point is not likely to be an important predictor. Both these tools provide local explanations only, namely they are able to provide insight for the outcome relative to specific observations, rather than giving a global rationale of the model functioning. However, LIME and SHAP are not the end of the story - there are a number of other methods that can be used, and the field of explainable AI (XAI) is rapidly developing. For more information, an excellent reference for current trends in Explainable AI is (Arrieta et al., 2020).

Explainability in AI will become increasingly more important as it enters

more and more facets of our daily lives. In particular, explainability provides transparency, promotes fairness and aids auditability. Without explainability methods, these three goals will be far harder to achieve.

Whilst naturally engineers and regulators will come into their roles with a certain set of skills, we argue that each group needs this set to be augmented to reason about ethical AI effectively. Engineers may understand how to efficiently train a model and increase its accuracy; however, this is only one side of the story - accuracy can be a poor metric to measure performance with[1], and the model may be actively discriminatory, for instance. It is crucial that engineers understand the statistical implications of what they are doing. Moreover, due to the issues of bias raised earlier in the document, we believe engineers should receive specific training in ethical AI and how to implement it. The exact training they should receive is up for debate, but should align with the material covered in (High-Level Expert Group on Artificial Intelligence, 2019) and should cover considerations such as fairness, accountability and transparency (Lepri, 2018). Second, we believe that given the complicated, fast-moving nature of AI as a technology, regulators should receive some sort of basic technical training, so they can appropriately judge and evaluate the systems they need to regulate. Whilst this training does not need to be as detailed as an engineer's training, it should be comprehensive enough that they are aware of the ethical issues that may arise during an engineer's implementation of a system.

Finally, we take inspiration from existing regulatory methods. For instance, in Norway, the data protection authority, Datatilsynet, has established a regulatory sandbox for developing AI solutions (IAPP, 2020), where industry works closely with regulators and policymakers to shape policy, while retaining the capacity to rapidly innovate. We believe this helps strike a good balance between innovation and regulation, allowing firms to move fast, whilst still remaining safe. For new applications and use cases, AI sandboxes offer a natural route for exploring these ideas, whilst still offering the potential for aggressive expansion. Another idea is to use escrow arrangements: while there is a growing practice within academia that the code for machine learning models is released along with the scientific paper, in a commercial environment, this may not be in a firm's interest. Thus, to provide assurances that an algorithm has not been tampered with, an escrow service could be used to hold a copy of the original dataset and algorithm and be verified by a third party service. If there were any doubts as to whether the original algorithm had been tampered with, the escrow service could feed an input into their version of the algorithm to show it produces the same results. This helps build trust in commercial ML systems, whilst still protecting corporate intellectual property.

1 Suppose we want to determine if a user has a fraudulent transaction on their account. Most transactions on a given person's account are legitimate, so if regardless of the data, we classify every transaction as non-fraudulent, then this model will have a high accuracy, but is effectively useless for the purpose of identifying fraud.

Recommendations

- A recommendation could be to embed certain capabilities or features in the AI system to allow it to be self-explanatory. In this way, the system would be able to tell the regulator, through specific mechanisms, why a certain decision has occurred in order to trace such decision;
- Regulators and engineers should be trained in explainability, and should ensure that appropriate methods are applied before deploying a system into production. Such methods are new and still in development but look very promising. We recommend that regulators in particular should be continuously informed by the latest developments in this field (for example, by regularly attending relevant conferences);
- The best way to ensure AI firms come up with models that are compliant is through them working with regulators as early as possible. The sandbox scheme which began in the UK with Fintech and is now being extended in some countries to specifically focus on AI (e.g. Norwegian AI Sandbox that began few months ago is a good example) is particularly efficient in achieving this goal. We highly recommend expansion of this scheme;
- Institutions should implement measures to ensure explainability of their AI/ML systems from the design phase. Even when full transparency cannot be achieved due to the intrinsic nature of the algorithm employed (e.g. deep neural networks), steps can be taken to identify and isolate in a human understandable format the main factors contributing to the final decision;
- Thoroughly document the development process leading to construction of AI/ML model since the design phase with assumptions and choices made at each step, for example:
    - » Document the blueprint of the data preparation flow used to build the input features, including transformations applied on the raw data (e.g. normalization, dimensionality reduction, exclusion of correlated features, aggregation, etc.), and the analysis performed to check the features' importance. Indeed, the information about which are the main input features influencing the model is already providing a first basic explanation of the model behaviour;
    - » Document the algorithms assessed and the comparative results justifying the selection made, including the analysis done on the model to ensure it is fit, the validation methodology employed and the results obtained. Indeed, different algorithms have different degrees of inherent complexity and the choice of the algorithm directly influences the model explainability;
    - » Document the metrics used to monitor model performance and promptly identify deviations.

- Keep the model as simple as possible: very often a simple model with an appropriate data pre-processing is as efficient and accurate as a very complicated one. This has the two-fold advantage of having a more transparent model and of fostering a deeper understanding of the dataset by the developers;
- When appropriate, embed interpreters into the model design to ensure traceability of the main internal steps leading to the final prediction:
    - » According to the criticality of the underlying use case and the need for transparency, evaluate the possibility to implement explainability techniques;
    - » Finally, ensure that the solution implemented in production is auditable, since audit logs can help to understand how data is processed (CSSF, 2018).

3.4. Accountability

Regulation is fundamentally a human activity. Regulators are humans - appointed to posts in organisations that operate on the basis of independent, non-political decision-taking and that are answerable to the executive or legislative arms of government as dictated by their governing law. In fulfilling their responsibilities, regulators interact with individuals in banking and finance institutions on a basis of trust, fortified by confidence that those individuals are "fit and proper" directors and controllers of their institutions, meeting tests of probity, competence, and financial standing. However, many B&F may rely on AI to aid decisions and operate their firms, the rationale for those decisions and operations must ultimately be explicable in assurances of regulatory compliance, i.e. assurances, expressed in non-AI language, that human beings (directors and controllers of firms) can provide to other human beings (prudential regulators). Thus, we believe it is important that there is a clear chain of responsibility and accountability whenever AI is used within an institution – if an algorithm is making decisions on behalf of a firm, then there should be someone who is accountable for its behaviour. This aids auditability and provides a first point of contact should any queries regarding the system arise.

Another option for helping ensure accountability is by introducing a firm-specific framework for how AI should be used internally. Some firms have experimented with this idea – one example is that of the Fair Banking Framework (Castelnovo et al., 2020) – but just like sets of principles for ethical AI, there is no universally adopted standard yet.

Recommendations

- A clear chain of accountability should be established wherever AI is deployed, assigning a person responsible to each instance of an algorithm used within a firm. All executives overseeing the departments where these models are used should receive training so they understand the implications of their deployment;
- Institutions should assume clear responsibility and accountability for the actions

and decisions taken by automated AI systems and processes. Ultimate responsibility should lie with the senior management of the institution which integrates the AI logic into its business processes. Whenever off-the shelf packages are acquired, clear liability provisions should be defined at contractual level. Furthermore, clear roles and responsibilities should be defined along all the AI lifecycle, including the development and operations activities, to ensure continuous engagement and accountability (CSSF, 2018);

- Introduce an internal code-of-conduct or framework for implementing AI within the firm

3.5 Human oversight

As we stressed before, it is crucial we view AI technologies as a set of tools to be used by humans, rather than as solutions to replace them. As such, the human element should be regarded as a crucial part of any AI production deployment, rather than as an afterthought. As elucidated in (High-Level Expert Group on Artificial Intelligence, 2019), which emphasises that "AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights, and allow for human oversight", it is important that AI systems do not run unchecked or unsupervised. Acknowledging human roles in AI entails a much bigger role for individuals than that of just a learner or supervisor. Individuals are now decision makers, who can interact, design, interrogate, and delegate work to the AI system, just like managers in an institution. Individuals should be able to define expectations, decide roles and decision rules.

A key example to explore is that of the 2010 Flash Crash. On 6 May 2010, starting at around 14:32 EDT, the Dow lost 9% of its value. Roughly forty minutes later, it regained 6 percentage points. Initially, the cause of this crash was debated – an early suspected culprit was a so-called "fat-finger trade", and there was even concern that some sort of cyberattack may have occurred. However, it later emerged that high-frequency, automated trading had caused a large bulk of the mass selling that occurred. Debatably, if humans had been more "in-the-loop" on large orders, such a crash would never have occurred. To avoid similar mistakes when using AI in B&F, it is crucial that we do not entirely defer to algorithms when making important decisions and ensure that appropriate checkpoints and human intervention/control or supervision mechanisms are in place.

Recommendations

- Human oversight helps ensure that an AI system does not undermine human autonomy or cause other adverse effects. The objective of trustworthy, ethical

and human-centric AI can only be achieved by ensuring an appropriate involvement by human beings in relation to high-risk AI applications.

- Even though the AI applications considered in this report for a specific legal regime are all considered high-risk, the appropriate type and degree of human oversight may vary from one case to another. It shall depend in particular on the intended use of the systems and the effects that the use could have for affected citizens and legal entities. For instance, human oversight could have the following, non-exhaustive, manifestations:

    - » the 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);
    - » the 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) (European Commission. 2020);
- 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 in any given condition from the preceding vehicle).

4. Final recommendations

Before summarising and collating our recommendations, it may seem there is nothing but innate risk in the use of AI. However, AI solutions can actually reduce a firm's exposure to risk. As established in the previous section, when AI technologies are used to automate business processes, their usage needs be regulated – as such, one idea is to actually use AI in these regulatory procedures themselves. This concept is known as Regulatory Technology, or RegTech. Whilst RegTech doesn't necessarily need to use AI, its adoption in recent years has certainly been powered by it. This direction offers a great deal of promise in relieving some of the burden traditionally associated with regulatory compliance.

RegTech

Regulatory Technology, or RegTech, is a burgeoning field which uses technology to help satisfy the regulatory requirements a firm might face. As Finance is already one of the most highly regulated industries, any tools which either partially or fully automate the work associated with regulatory compliance have the potential to provide a huge benefit for a firm.

Deloitte (Hugé et al., 2020) characterise the RegTech industry by dividing it into five broad categories:
1) Regulatory Reporting – automatically collecting, collating and publishing the data required for firm's regulatory reporting requirements;
2) Risk Management – monitoring and mitigating risk exposure, both in the sense of commercial risk and regulatory risk;
3) Identity Management & Control – automating KYC and AML processes, as well as customer due diligence and anti-fraud procedures;
4) Compliance – tracking adherence to current, as well as future, regulatory requirements;
5) Transaction Monitoring – providing real time monitoring of transactions, often through the use of Blockchain technologies.

Given the increasing amount of regulation in recent years (for instance, the measures introduced after the 2008 financial crisis, and GDPR coming into force in 2018), the costs associated with compliance have increased. Moreover, regulatory adherence has historically been a time-intensive activity, reducing the capacity of the workforce. However, a key theme within the above five categories is that of automating existing processes – as such, RegTech offers a new approach to relieving the financial and temporal burdens of regulatory requirements, allowing firms to be more agile, whilst still adhering to the standards required of them.

a) Recommendations for Regulators

The best way to ensure AI firms come up with models that are compliant is through working with regulators as early as possible. Schemes such as the sandbox scheme which began in the UK with FinTech and is now being extended in some countries to specifically focus on AI (e.g. the Norwegian AI Sandbox) is particularly efficient in achieving this goal. We highly recommend expansion of such schemes. We understand the EU is already considering an EU-wide sandbox and we believe that for many AI applications such approach would be very helpful.

In cases where gaps still remain between the tech and the regulators' understanding or trust in its various risk management arrangement, we recommend copies of the code are kept by the regulators in escrow so that it can be run on data independently to check if agreed risk parameters have been interfered with. This is common practice in algorithmic trading in hedge fund management and is known to be effective.

We recommend that regulators hire people with technical background and provide regular training to all practitioners involved with regulating banking and finance AI application. In this training, we recommend the inclusion of the following topics in the context of B&F: statistical verification (e.g. underfitting and overfitting and the risk associated with explaining the past rather than predicting the future), bias and fairness training, and explainability. There already exist some programmes such as the Oxford Algorithmic Trading programme where such topics are explained clearly to non–academic audiences. We also recommend that especially with new, emerging topics (such as explainability, for instance), regulators should receive regular updates on the latest AI tech advancements and what is emerging in research and academia.

b) Recommendations for Industry

Wherever AI is deployed, a clear chain of accountability should be established, assigning a person responsible to each instance of an algorithm used within a firm. All executives overseeing the departments where these models are used should receive training so they understand the implications of their deployment. Additionally, engineers and managers should receive training pertaining to ethical AI and what it means for their systems. Awareness of the relevant regulation is important at every stage of development of AI systems and can prevent problems later on. Engineers should be encouraged to take on training in statistical analysis of the models and the datasets – in particular statistical verification and fitting (and the risks of overfitting).

Furthermore, traditional software testing should remain a critical part of any AI deployment, and new ideas from neural network verification should also be integrated into the process. Engineers should be also aware of adversarial attack vectors, and receive training in the latest news in this area and learn how to mitigate such risks (as much as possible).

Finally, engineers should become aware of risks such as how models "leak" data and receive training to better understand these risks and how to mitigate them. Instead of trying to entirely eliminate bias, we should learn to manage it instead. This includes being able to identify potential for bias in the models and datasets and understanding the types of algorithms that can mitigate its effect.

References

Amazon Web Services. 2020. "AWS AI | Coinbase." Amazon Web Services, Inc. Accessed November 16, 2020. https://aws.amazon.com/machine-learning/customers/innovators/coinbase/.

Arrieta, Alejandro Barredo, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." Inf. Fusion 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 50 (November 2019), 24 pages. https://doi.org/10.1145/3359152

Branwen, Gwern. 2011. "The Neural Net Tank Urban Legend." www.gwern.net. https://www.gwern.net/Tanks.

Carlini, Nicholas, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks." In 28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019, edited by Nadia Heninger and Patrick Traynor, 267–84. USENIX Association. https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

Chatfield, Daniel. 2017. "Fighting Fraud with Machine Learning." Monzo. https://monzo.com/blog/2017/02/03/fighting-fraud-with-machine-learning.

Castelnovo, Alessandro, Crupi, Riccardo, Greco, Greta, Del Gamba, Giulia, Naseer, Aisha, Regoli, Daniele, and San Miguel Gonzalez, Beatriz. 2020. "BeFair: Addressing Fairness in the Banking sector." Proceedings of the IEEE International Workshop on Fair and Interpretable Learning Algorithms (to appear).

Commission de Surveillance du Secteur Financier (CSSF). 2018. "Artificial Intelligence: opportunities, risks and recommendations for the financial sector". https://www.cssf.lu/wp-content/uploads/files/Publications/Rapports_ponctuels/CSSF_White_Paper_Artificial_Intelligence_201218.pdf

Council of Europe, Committee of experts on Internet MSI-NET. 2017. "Study on the human rights dimensions of automated data processing techniques and possible regulatory implications".

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." CoRR abs/1810.04805. http://arxiv.org/abs/1810.04805.

The Economist. 2018. "AI, Radiology and the Future of Work." The Economist. https://www.economist.com/leaders/2018/06/07/ai-radiology-and-the-future-of-work.

Bank of England, and Financial Conduct Authority. 2019. "Machine Learning in UK Financial Services." https://www.fca.org.uk/publication/research/research-note-on-machine-learning-in-uk-financial-services.pdf.

Equifax. 2017. "Cybersecurity Incident & Important Consumer Information | Equifax." 2017 Cybersecurity Incident & Important Consumer Information. https://www.equifaxsecurity2017.com/.

European Commission. 2020. "White Paper on On Artificial Intelligence - A European approach to excellence and trust". https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

European Parliament. 2020. Framework of ethical aspects of artificial intelligence, robotics and related technologies. https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_EN.pdf.

Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. 2015. "Model Inversion Attacks That Exploit

Confidence Information and Basic Countermeasures." In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015, edited by Indrajit Ray, Ninghui Li, and Christopher Kruegel, 1322–33. ACM. https://doi.org/10.1145/2810103.2813677.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2015. "Explaining and Harnessing Adversarial Examples." In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, edited by Yoshua Bengio and Yann LeCun. http://arxiv.org/abs/1412.6572.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. MIT Press.

Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2020. Toward fairness in AI for people with disabilities SBG@a research roadmap. SIGACCESS Access. Comput., 125, Article 2, 1 pages. https://doi.org/10.1145/3386296.3386298

Hendrycks, Dan, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2019. "Natural Adversarial Examples." CoRR abs/1907.07174. http://arxiv.org/abs/1907.07174.

High-Level Expert Group on Artificial Intelligence. 2019. "Ethics Guidelines for Trustworthy AI." European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation.

Hofstadter, Douglas R. 1979. Gödel, Escher, Bach : An Eternal Golden Braid. Harvester Studies in Cognitive Science ; 12. Hassocks: Harvester.

Hugé, F.-K., Laurent, P., Ramos, S. and Berliner, L. (2020). RegTech Universe. [online] Deloitte Luxembourg. Available at: https://www2.deloitte.com/lu/en/pages/technology/articles/regtech-companies-compliance.html. Accessed 28 Nov. 2020.

IAPP. 2020. "Norwegian DPA Creating Regulatory Sandbox for AI." https://iapp.org/news/a/norwegian-dpa-creating-regulatory-sandbox-for-ai/.

International Regulatory Strategy Group, and Accenture. 2019. "Towards an AI-Powered UK: UK-Based Financial and Related Professional Services." https://www.irsg.co.uk/publications/irsg-report-towards-an-ai-powered-uk-uk-based-financial-and-related-professional-services/.

Ji, Zhanglong, Zachary Chase Lipton, and Charles Elkan. 2014. "Differential Privacy and Machine Learning: A Survey and Review." CoRR abs/1412.7584. http://arxiv.org/abs/1412.7584.

Khandani, Amir E., Adlar J. Kim, and Andrew Lo. 2010. "Consumer Credit-Risk Models via Machine-Learning Algorithms." Journal of Banking & Finance 34 (11): 2767–87. https://EconPapers.repec.org/RePEc:eee:jbfina:v:34:y:2010:i:11:p:2767-2787.

Kleinberg, Jon, Sendhil Mullainathan, Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." Proceedings of Innovations in Theoretical Computer Science (ITCS). https://arxiv.org/abs/1609.05807

Knight, Will. 2019. "The Apple Card Didn't 'see' Gender—and That's the Problem." Wired, November. https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/.

Krebs, Brian. 2019a. "First American Financial Corp. Leaked Hundreds of Millions of Title Insurance Records — Krebs on Security." Krebsonsecurity.com. https://krebsonsecurity.com/2019/05/first-american-financial-corp-leaked-hundreds-of-millions-of-title-insurance-records/.

———. 2019b. "What We Can Learn from the Capital One Hack — Krebs on Security." Krebsonsecurity.com. https://krebsonsecurity.com/2019/08/what-we-can-learn-from-the-capital-one-hack/.

Kusner, Matt J., Joshua R. Loftus, Chris Russell, Ricardo Silva. 2017. Counterfactual Fairness. Advances in Neural Information Processing Systems 30 (NIPS 2017). https://arxiv.org/abs/1703.06856.

Lanier, Jaron. 2019. Ten Arguments for Deleting Your Social Media Accounts Right Now. London.

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. "Fair, Transparent, and Accountable Algorithmic Decision-Making Processes." Philosophy & Technology 31 (4): 611–27.

Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. "Towards Deep Learning Models Resistant to Adversarial Attacks." CoRR abs/1706.06083. http://arxiv.org/abs/1706.06083.

Mehrabi, Ninareh, et al. 2019. "A survey on bias and fairness in machine learning." arXiv preprint arXiv:1908.09635.

Mittelstadt, Brent. 2019. Principles alone cannot guarantee ethical AI. Nat Mach Intell 1, 501–507. https://doi.org/10.1038/s42256-019-0114-4

Moskowitz, Tobias J. 2016. "Momentum Crashes." www.aqr.com. https://www.aqr.com/Insights/Research/Journal-Article/Momentum-Crashes.

Naughton, John. 2019. "Can the Planet Really Afford the Exorbitant Power Demands of Machine Learning? | John Naughton." The Guardian; The Guardian. https://www.theguardian.com/commentisfree/2019/nov/16/can-planet-afford-exorbitant-power-demands-of-machine-learning.

Noya, Eloi. 2019. "The Fintech Revolution: Who Are the New Competitors in Banking?" Forbes, July. https://www.forbes.com/sites/esade/2019/07/30/the-fintech-revolution-who-are-the-new-competitors-in-banking/.

OECD. 2017. "Algorithms and Collusion: Competition Policy in the Digital Age." http://www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm.

Pinsent Masons, and Innovate Finance. 2019. "AI in Financial Services Impact on the Customer."

Pound, Jesse. 2019. CNBC. https://www.cnbc.com/2019/12/24/global-stock-markets-gained-17-trillion-in-value-in-2019.html.

Pedreschi, Dino, Ruggieri, Salvatore, and Turini, Franco. 2008. Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08). Association for Computing Machinery, New York, NY, USA, 560–568. DOI:https://doi.org/10.1145/1401890.1401959

Reuters. 2016. "Wall Street Watchdogs Turn to Artificial Intelligence." Fortune. https://fortune.com/2016/10/25/how-artificial-intelligence-could-catch-stock-market-cheaters/.

Rhodes, Chris. 2019. "Financial Services: Contribution to the UK Economy." House of Commons Library. https://commonslibrary.parliament.uk/research-briefings/sn06193/.

Robert, L. P., Gaurav, B., and Lütge, C. 2020. ICIS 2019 SIGHCI Workshop Panel Report: Human–Computer Interaction Challenges and Opportunities for Fair, Trustworthy and Ethical Artificial Intelligence. AIS Transactions on Human-Computer Interaction, 12(2), pp. 96-108.

Saxena, Nripsuta Ani, Huang, Karen, DeFilippis, Evan, Radanovic, Goran, Parkes, David C., and Liu, Yang. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (Honolulu, HI, USA) (AIES '19). Association for Computing Machinery, New York, NY, USA, 99–106. https://doi.org/10.1145/3306618.3314248

Sharif, Mahmood, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. 2016. "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition." In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016, edited by Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and

Shai Halevi, 1528–40. ACM. https://doi.org/10.1145/2976749.2978392.

Sitawarin, Chawin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. 2018. "DARTS: Deceiving Autonomous Cars with Toxic Signs." CoRR abs/1802.06430. http://arxiv.org/abs/1802.06430.

Song, Congzheng, Thomas Ristenpart, and Vitaly Shmatikov. 2017. "Machine Learning Models That Remember Too Much." CoRR abs/1709.07886. http://arxiv.org/abs/1709.07886.

Stripe. 2020. "Stripe Radar: Fraud Prevention for Credit Cards & Payments." stripe.com. Accessed November 16, 2020. https://stripe.com/gb/radar.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. "Intriguing Properties of Neural Networks." In 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, edited by Yoshua Bengio and Yann LeCun. http://arxiv.org/abs/1312.6199.

Thanendran, Abhi. 2018. "How We Use Machine Learning to Protect You from Fraud | Revolut." Revolut Blog. https://blog.revolut.com/how-we-use-machine-learning-to-protect-you-from-fraud/.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, edited by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Verma, Sahil and Rubin Julia. 2018. "Fairness definitions explained." IEEE/ACM

International Workshop on Software Fairness (FairWare).

Vulkan, Nir. 2019a. Oxford Programme on FinTech

Vulkan, Nir. 2019b. Oxford FinTech MBA elective, Said Business School, Oxford.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, edited by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, 5754–64. http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.

Zhang, Yukun, and Longsheng Zhou. 2019. Fairness Assessment for Artificial Intelligence in Financial Industry. arXiv preprint arXiv:1912.07211.

3
ENERGY

Authors

Nicolae Lucian Mihet
Professor in Energy Technology, Faculty of Engineering, Oestfold University College, Norway

Afzal S. Siddiqui
Professor of Energy Economics in the Department of Statistical Science, UCL, UK

Fausto Pedro García Márquez
Full Professor at Castilla-La Mancha University, Spain

Rónán Kennedy
Lecturer in Law, School of Law, National University of Ireland Galway, Irelands

Sergio Saponara
Professor of Electronics, Department of Information Engineering, Pisa University, Italy

Chapter 1. Introduction

1. Background and Overview

The European Commission defines AI as "systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g., voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g., advanced robots, autonomous cars, drones or Internet of Things applications)" [1].
Figure 1 shows the main AI strategies and some national plans between 2017 and 2019 [2].

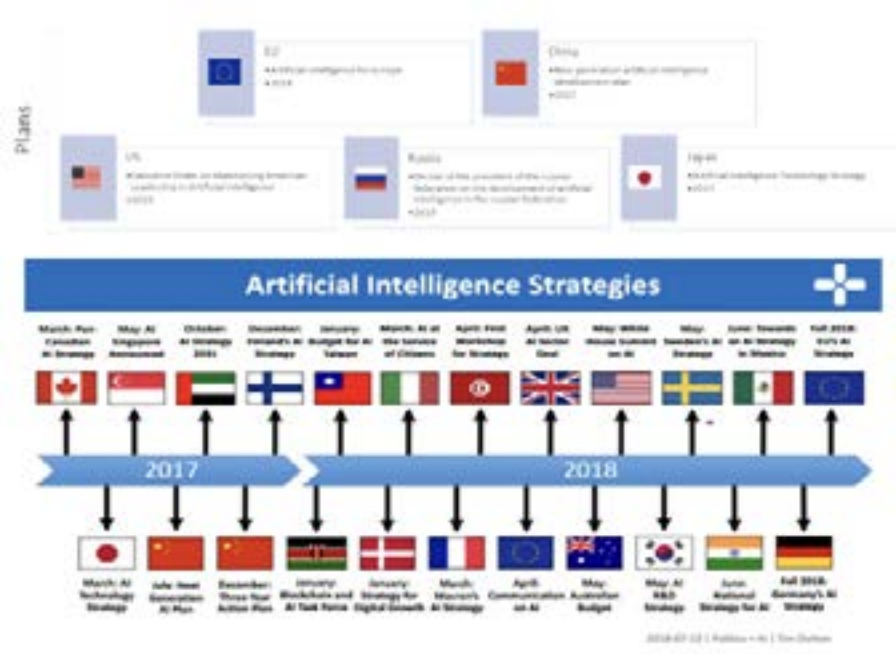


Fig. 1. AI strategies by countries.

1.2. Aim and Objectives

The aim of this report is to provide comprehensive guidelines for the development of AI for the energy industry sector with practical recommendations based on fundamental rights and the ethical principles of the seven key requirements. These are applicable to different stakeholders (developers, deployers, end users, and society) engaged in the life cycle of AI systems, having equal importance but different roles to play in ensuring that the requirements are met.

Trustworthy AI can represent a great opportunity to support the mitigation of pressing challenges facing society such as climate change. While tackling climate change should be a top priority for policymakers across the world, the digital transformation and trustworthy AI have a great potential to reduce human impact on the environment and enable the efficient and effective use of energy and natural resources [3]. For instance, digital transformation and trustworthy AI can be coupled with big data analysis in order to detect energy needs more accurately, which will provide more efficient energy infrastructure and consumption.

The concrete objectives for the energy industry are to create an ecosystem of trust and excellence along the entire value chain, to create the right incentives for accelerating the adoption of solutions-based AI, including by SMEs, and to consider the key elements of a future regulatory framework for AI in Europe to ensure compliance with EU rules [4].

Also, combining its technological and industrial strengths with a high-quality **digital infrastructure** and a regulatory framework based on its fundamental values, the EU can become a global leader in innovation of the data economy and its application in the energy sector. This brings the benefits of AI technology to *EU citizens, business development*, and *services of public interest.*

1.3. Report Outlines

Chapter 1 gives a short introduction to this report highlighting its aim and objectives. Chapter 2 provides comprehensive guidelines for the application of the seven key requirements and their impact on various energy industry applications while Chapter 3 suggests concrete and practical steps that the energy sector must take to be compliant with the seven key requirements. In Chapter 4, some conclusions are drawn as well as practical recommendations for AI adoption in the energy industry.
The correspondence and connection between chapters and topics in this report is highlighted in Fig. 2.

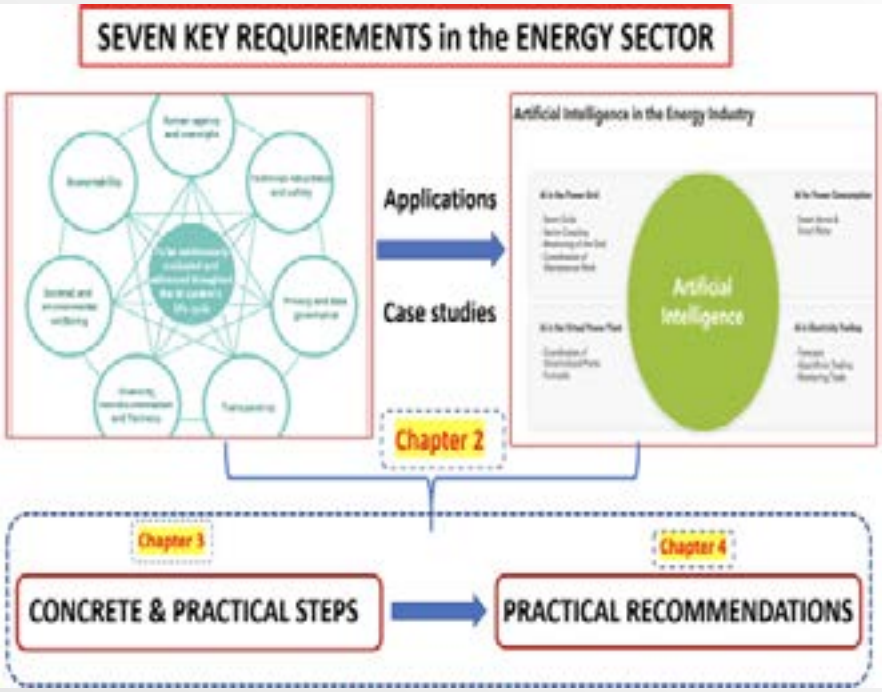


Fig. 2. A block diagram with report outline.

Chapter 2.
How the Seven Key Requirements Impact the Energy Sector

2.1. Concrete Objectives and Key Factors Considering the Impact

In order to be prepared for Industry 4.0, we have to envisage how AI, big data, and machine learning can transform the way we work and which tools are available or suitable to work alongside robots and machines in the energy sector. We also need to optimize the work that is necessary so that the interaction between humans and robots in shaping the future workplace is tailored to the needs of society as a whole.

The aim of the framework, in partnership between the private and public sectors, is to:

-- **Create an ecosystem of excellence** along the entire value chain

-- **Conceive the right incentives for accelerating the adoption of solutions-based AI**, including by SMEs (digital innovation hubs should provide support to SMEs to understand and adopt AI)

-- **Specify the key elements** of a future regulatory framework for AI in Europe **that will create a unique ecosystem of trust to ensure compliance with EU rules.**

-- Combine technological and industrial strengths with a **high-quality digital platform**/infrastructure

2.2. The Seven Key Requirements (K1-K7)

Based on the fundamental rights and ethical principles, **the seven key requirements** that AI systems should meet in order to be trustworthy are listed below:

> **K1. Human agency and oversight – including fundamental rights, human agency, and human oversight**

The most important fundamental rights that companies within the energy sector should consider are:
-- Dignity and non-discrimination;
-- Personal data and privacy protection;
These are considered in detail later in this document.

Human agency:

Uses of AI in the energy industry must take account of the fact that individuals and groups will have divergent but still legitimate goals, and allow the possibility of exercising choice. AI must enable better, more informed human decision-making, not take away control. In particular, the development of large-scale and long-term infrastructure, whether physical or virtual, must consider the need to preserve flexibility for the future, so that as political preferences, scientific understanding, and social conditions change, new approaches can be implemented with as little difficulty and expenses as possible. There must always be scope for alternative conceptions and configurations of the marketplace, such as decentralized, community-owned, or not-for-profit approaches. Preserving the possibility for innovation is crucial. In addition, individuals and businesses must be able to make free and informed choices about how to manage their energy use and expenditure. Transparency in the ways in which AI, big data, and machine learning are being applied to manage demand, determine tariffs, and smooth charges are essential foundations for this, particularly for those suffering from energy or financial poverty.

Human oversight:

Although AI provides a mechanism for making decisions at an otherwise impossible pace and scale, human oversight must be maintained. **Renewable energy systems (RES) must be designed so that unusual events or fluctuations are brought to the attention of operators as quickly as possible, and intervention is always possible, so that runaway algorithms or cascading failures are avoided. Human oversight is also an important mechanism for identifying security breaches and risks, and, therefore, systems should be both auditable and regularly audited**. This is discussed further in the next section on technical robustness and safety. Machine learning models must also be regularly checked and verified to ensure that they have not developed rules that are socially harmful, biased, or otherwise undesirable.

K2. Technical robustness and safety – including resilience to attack and security, fall-back plan and general safety, accuracy, reliability and reproducibility

Having access to stable and reliable energy supply is among the key rights of citizens. Thus, technical safety and security of energy systems, including resilience to attacks, are key elements for Europe.

Moreover, the energy supply and the entire energy system (i.e., energy production, trading, distribution, storage, and utilization) are part of critical European infrastructure affecting European industry, logistics, and mobility (of both people and goods).
To summarize, therefore, **AI technology should contribute to the improvement of safety and technical robustness, as well as the identification and mitigation of attacks.** To this end some security and safety aspects are discussed below and some key technologies, enabled by AI, are highlighted.

Technical robustness and safety

In terms of technical robustness and safety, all parts of the energy system (production, trading, distribution, storage, and utilization) should be supported by predictive health maintenance technologies where AI can give its key contribution. By analyzing both physical data or results from model estimations, an AI based prediction health manager can reduce the risks of faults and ageing degradation and hence can reduce the risk of accidents or denial of service.

Thanks to predictive management, all maintenance operations can be scheduled during ordinary maintenance steps, thereby reducing severity and the probability of faults.

Predictions based on AI techniques can be used not only for energy production, storage and distribution, but also for energy trading, e.g., to find the right compromise between energy production-selling price-user consumption in a free but transparent energy market.

With the increasing decentralization and digitalization of the power grid/ smart grid, it is becoming more difficult to manage the large number of grid participants and to keep the grid in balance. This requires evaluating and analyzing a large amount of data. AI helps to process, evaluate, analyze, and control these data as quickly and efficiently as possible.

Resilience to attack and security

Differing from faults or aging degradation, failures due to cyberattacks may be difficult to detect by basic analysis of physical acquired data. To overcome this obstacle, the

energy system should also be equipped with anomaly detection tools. Exploiting AI techniques, fingerprinting and anomaly detection methods aim to detect anomalies in the behavior of a network or of a software (SW)-defined control system. Fingerprinting and anomaly-detection methods also aim to classify if there is a failure due to aging or random fault or a failure due to malicious software.

The security problem is exacerbated by the increased connectivity of the energy system (production plants, energy storage stations, distribution grid and recharging stations) since due to the smart grid paradigm and the V2G (vehicle to grid) approach the range of possible attacks is increasing. Moreover, the capability of remote update of the SW of an energy control unit has the drawback of increasing the possibility of malicious software's taking control of an energy subsystem.

Fingerprinting and anomaly detection is the first step to classify faults and cyber-attacks correctly thereby providing a fall-back plan to ensure safety of the energy system.

AI-based fingerprinting and anomaly detection can be used not only for energy production, storage, and distribution plants, but also to detect anomalies during energy trading sessions for a transparent energy market.

Fall-back plan and general safety

As discussed above, AI can enable predictive health maintenance and anomaly detection and fingerprinting techniques to be adopted in energy systems. However, to ensure that energy systems are resilient and safe, these prediction and detection capabilities should be complemented by the presence of fall-back energy plans in case something goes wrong. Here also, AI can be useful to define the optimal level of redundancy finding a trade-off among complexity, cost, and recovery performance of the energy back-up solution.

Accuracy

The proposed solutions for AI-based predictive health maintenance and anomaly detection above should be tested and validated to ensure accuracy, minimizing false alarms or missed detections. Indeed, in industrialized countries, such as the EU Member States, the power supply is already quite reliable, and hence any change to a new one (perhaps improved in terms of affordability or green sustainability) using AI-prediction techniques must be accurate. Unintended consequences should be avoided, as well as both false alarms or missed detections. Similar accuracy of AI-based predictions are key for their success in the energy market (e.g., electricity trading).

Reliability and reproducibility

Reliable energy solutions must be also reproducible to allow scaling to different contexts (industrial, domestic, mobility, and transport system) and rules (e.g., different national regulatory frameworks). Moreover, reproducibility is a key element to allow system validation and testing and, through large-scale production, to allow for affordability.

K3. Privacy and data governance – including respect for privacy, quality, and integrity of data, and access to data

The volume and the way in which data are stored and processed will change dramatically in the next five years. As discussed in the previous section, a key issue for energy system sustainability and affordability is the capability to profile and predict the energy needs of the energy users (citizens, but also industries, logistics, and mobility systems). One of the aims of using AI technologies is to optimize the way energy is used, shared, produced, and traded. All of these aspects require the development of distributed intelligence needing data analytics and exchange of a large volume of data. However, the level of profiling possible can enable the identification of very private details of an individual's lifestyle, routines, and habits, which could be extrapolated to make inferences about their health.

Hence, thanks also to AI, privacy of sensitive data should be ensured and data governance should be put in place.

To give the energy industry and, in particular, end-use consumers more confidence in AI (e.g., in relation to smart-home technologies, smart meters, or optimization of charging infrastructure in smart grid applications for electrified mobility), it must be clearly communicated how the data are used and by whom, while data security must be guaranteed. Therefore, the European Commission has developed four ethical principles for AI: AI should respect human autonomy, avoid social harm, be fair, and be explainable [4].

In addition to following ethical guidelines, it is essential that AI in the energy sector comply fully with European law, particularly the General Data Protection Regulation (GDPR). It should also apply the Data Protection Impact Assessment Template prepared by the Commission's. Smart Grid Task Force [5], and follows the Commission's recommendation (2012/148/EU) on the roll-out of smart metering systems, which underlines the importance of taking all reasonable steps to ensure that data cannot be de-anonymized, incorporating data protection by design and data protection by default into the methodologies used for the development of smart grids, and designing in data security from an early stage.

Data governance will be important in order to ensure compliance with data protection law and to maximize the potential use and re-use of energy data for modeling, machine learning, and other applications in the future. Energy systems must have mechanisms in place to track the nature of collected data, to determine the lawful basis for their processing, and to manage sharing, either within an organization or with third parties. Care must be taken to ensure that future processing is in line with the original purposes for which data were collected. Furthermore, if the lawful basis was consent, than it should be ensured that future processing is within the scope of that consent and that there is some mechanism for those who have withdrawn consent to have their data removed from all copies of a dataset, particularly if they have been externally shared. Individuals must be able to vindicate their rights of access, erasure, and data portability, in line with the recommendations of the European Smart Grids Task Force Expert Group 1 report on "My Energy Data" [6]. This may prevent the development of AI applications that are technically possible but cannot comply with the law. The early application of data protection by default and by design methodologies to new ideas will prevent wasting resources on projects that cannot be legally completed.

The energy supply and the entire energy system are part of critical infrastructure. This is why cybersecurity is becoming increasingly important in order to protect the highly networked power grid/smart grid from attacks and data theft from the outside. Consequently, there should be clear security requirements for participants in the electricity market to ensure data confidentiality, integrity, and authentication along the whole chain of data acquisition, processing, and storage. This can be achieved by defining best practices to be followed for security of data taking note of the different roles: users, production industries, companies working as service providers, and institutions (e.g., standardization bodies, regulatory authorities, and governments).

Data protection and data security (especially mitigation of cyberattacks) are very sensitive points for the use of AI in the energy industry. AI can make an important contribution in the fight against cyberattacks by quickly checking large amount of data and thus detecting deviations. The Smart Grid Task Force Expert Group 2 report on cybersecurity [7], which proposes a Network Code on Cybersecurity for energy system operators, should be taken into consideration.
Toward this end, the AI4 People initiative can find synergies with the IEEE Global Initiative on Ethically Aligned Design for a Sustainable Planet [8]/). More details are available in Section 2.2/K7.

K4. Transparency – including traceability, explainability, and communication

As with other sectoral applications of AI, its acceptance will depend on its transparency. In the financial and legal sectors, AI is increasingly deployed with widespread ramifications for society, e.g., in terms of credit scoring and parole decisions. Just like other forms of automation, AI can create value by relieving human intervention in routine tasks that are based on processing increasingly large amounts of data. However, AI's mechanisms can appear opaque, thereby undermining its credibility. Thus, AI applications with societal consequences may require an auditing procedure that can trace the logic of its decision-making nexus. This is not dissimilar to financial auditing, which was necessitated by the advent of publicly held companies that issued stock. Such a process provided creditors and investors with a snapshot of the companies' financial architecture. Likewise, regulatory standards for AI audits could be established.

In the energy sector, these would involve testing AI on prototype use cases, e.g., energy trading or economic dispatch, that could be compared with output from conventional decision-making approaches. In a sense, regulators in the electricity industry already face this challenge due to deregulation as power companies use sophisticated offering strategies in order to maximize profit, which could be at odds with the maximization of social welfare. The task of monitoring markets could be daunting, but it is aided somewhat in the electricity industry due to physical network constraints and the need for energy balance in real time. **Hence, marshaling this existing expertise together with algorithmic advances in inverse optimization [9]9 could be used by auditors to unveil how AI applications function in the energy sector and to ensure that their use is in line with social objectives.**

K5. Diversity, non-discrimination, and fairness – including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

The diversity problem is about gender, race, and most fundamentally, about power and energy. It affects how AI companies work, which products get built, who they are designed to serve, and who benefits from their development.

Matching characteristics of the energy market to models of discrimination, we need to identify the necessary conditions for the license condition to have a positive effect for consumers, and to explore whether the policy has helped potentially vulnerable consumers. Non-discrimination law and data protection law are the main legal instruments that could protect the right to non-discrimination in the context of algorithmic decision-making.

Algorithmic decision-making and other types of AI can threaten human rights, such as the right to non-discrimination. In the private sector, algorithmic decision-making can have discriminatory effects. For instance, AI can be used by firms to select

employees, while targeted online advertising is largely driven by algorithmic decision-making. Such advertising is a profitable sector for some companies. For example, Facebook and Google make most of their money from online advertising. However, online advertising can have discriminatory effects. While decisions by algorithms can have discriminatory effects, algorithms are not inherently bad or discriminatory. Algorithms might still perform better than human decision-makers.

Non-discrimination law and data protection law are the most relevant legal instruments to fight illegal discrimination by algorithmic systems. If effectively enforced, then both legal instruments can help to protect people.

Biased training data, however, could lead to discriminatory decisions. The training data can be biased because they represent discriminatory human decisions. **In order to ensure non-discriminatory programming and functioning, the systems need to be trained and tested for unfair bias. Therefore, companies should test their algorithms for bias and discrimination and demonstrate that certain fairness standards are met [10]. The IEEE P7003 standard** for algorithmic bias considerations lays out relevant instructions for eliminating issues of negative bias when developing algorithms.

The European Economic and Social Committee (EESC) suggests that the EU should develop a certification for trustworthy AI applications, to be delivered by an independent body after testing the products for key requirements such as resilience, safety, and absence from prejudice, discrimination or bias [11].

K6. Societal and environmental wellbeing – including sustainability and environmental friendliness, social impact, society, and democracy

Given the focus of Europe on climate-change, all technologies like ICT and AI applied to energy systems should improve environmental sustainability and the affordability of green energy solutions for citizens.

Key initiatives to this end are the development and adoption of renewable energy sources (RES) (even at the small scale of rooftop photovoltaic or hybrid plants at single building level) and the possibility to create peer-to-peer community of energy prosumers (producers/consumers), to profile/predict their energy needs and to optimize the way energy is used, shared, produced, and traded. All of these aspects require the development of a distributed intelligence needing data analytics and a large exchange of data.

AI in power trading helps to improve forecasts (based on weather data, historical data, consumption profiles and mobility patterns) as well as to facilitate and speed-up the integration of renewables (green energy). Better forecasts also increase grid stability and, thus, security of supply. For example, wind power futures for Germany have existed since 2017 for hedging their risks [12].

AI systems could mitigate climate change and environmental degradation by contributing to positive solutions for critical resource usage and energy consumption. Consumers, intelligently connected in the electricity system, can contribute to a stable and green electricity grid. Smart home solutions and smart meters already exist but they are not yet widely used. Part of the obstacle is the lack of a business model, which leads to market failure. In essence, individual consumers on their own do not have the economic incentive to deploy energy-efficiency technologies. However, given the mechanism to pool consumption profiles and preferences, an aggregator using AI could profit from utilization of such demand-reduction and load-shifting potential at a scale that could facilitate the integration of both demand-side management and RES technologies.

Related to K4 above, deployment of AI by profit-maximizing entities may not be fully aligned with societal objectives such as welfare maximization, equity, and emission reduction. Besides the auditing procedure alluded to in K4 above, empowerment of citizen groups to harness AI could provide a formidable counterparty on the demand side. Thus, by enabling both sides of the energy market to function, it may be possible for AI to keep itself in check from putting suppliers' interests above those of consumers. **Regulatory authorities' use of AI would add another layer of intelligence that could anticipate the actions of market participants and tweak market designs to discourage deviations from societal objectives.** Underpinning legislation already exists in Article 16 (local energy communities) from the EU's clean energy package [13] . By requiring Member States to ensure non-discriminatory access to all markets through established or autonomously managed networks, it could provide an AI dimension to energy communities to flourish.

For consumers on low incomes and suffering from energy poverty, smart meters with dynamic pricing managed by AI systems may provide an opportunity to save money, but the difference that this makes is likely to be marginal, while much of their consumption cannot be moved to lower-price time periods. They are also less likely to be able to benefit from feed-in tariffs. The cost of prepayment should, therefore, be kept as low as possible, so that all have an opportunity to participate fully in new energy markets.

K7. Accountability – including auditability, minimization and reporting of negative impact, trade-offs, and redress

The EC defines accountability as "Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes.]
Auditability, which enables the assessment of algorithms, data, and design processes, plays a key role in this, especially in critical applications. Moreover, adequate and accessible redress should be ensured" [3]. The EC has announced the following requirements related to the accountability in AI together with the dataset:

- The use of AI-enabled products and services must be safe. They must meet EU safety rules (existing as well as possible complementary ones) and the standards set. These standards and EU safety rules are now changing worldwide, (see below), with new proposals to be completed and approved. The Standards Development Organizations (SDOs) develop IT standards using different models to address varying standardization needs.
- The datasets would be available upon request for any competent authorities, e.g., inspections.
- The process should protect confidential information (e.g., trade secrets).
- The datasets employed to train the AI systems need to consider all the scenarios required and to be sufficiently broad.
- The datasets, documentation, and records, would need to be retained over a sufficient time period (reasonable and limited) to guarantee the application of the aforementioned standards and EU safety rules.
- Competent authorities and affected parties should be informed about the limitations and capabilities of the AI systems, mainly about its purpose, functionality conditions.

Since people are at the center of the AI strategy set by the EC [2], in the case of any AI system which works without human interaction, the people affected by the AI system should be informed, with easily understandable, concise, and objective information. The appropriate form will depend on the particular context [4]. The EU data protection legislation considers this [14], but the EC is working on new rules and standards to guarantee it. By contrast, in cases where it is sufficiently clear that citizens are interacting with AI systems, this information does not need to be provided.

The competitiveness of European business and a high standard of safety would be reached by a suitable legal environment. The damage caused by AI systems generates an unclear liability scenario under the Product Liability Directive, and it is not clear also how to apply this Directive to certain defects due to AI systems [14]. The General Product Safety Directive [15] is the EU legal framework for product safety to provide

a high level of safety and health. This Directive must be considered together with the different systems of civil liability for damages caused by products or services.

The U.S. President issued an Executive Order in 2019 to "Ensure that technical standards reflect Federal priorities for innovation, public trust, and public confidence in systems that use AI technologies and develop international standards to promote and protect those priorities" [16]. It directs the Secretary of Commerce, through the National Institute of Standards and Technology (NIST), to issue "a plan for Federal engagement in the development of technical standards and related tools in support of reliable, robust, and trustworthy systems that use AI technologies" [17]. The AI standards will consider the following aspects:

- Concepts and terminology; Data and knowledge; Human interactions; Metrics; Networking; Performance testing and reporting methodology; Safety; Risk management; Trustworthiness (guidance and requirements for accuracy, explainability, resiliency, safety, reliability, objectivity, and security).

According to the plan, the tools/applications include, but are not limited to the following [17]:

- Datasets in standardized formats, including metadata for training, validation and testing of AI systems; Tools for capturing and representing knowledge, and reasoning in AI systems; Fully documented use cases that provide a range of data and information about specific applications of AI technologies and any standards or best practice guides used in making decisions about deployment of these applications; Testing methodologies to validate and evaluate AI technologies' performance; Metrics to quantifiably measure and characterize AI technologies; Benchmarks, evaluations, and challenge problems to drive innovation; AI testbeds; Tools for accountability and auditing.

Standards under the direct responsibility of ISO/IEC secretariat related to AI are described in Appendix I.

China has published the "New generation artificial intelligence development plan" [18]. According to point V. Guarantee measures, the strategy fixed the following issue:

- Establish an AI technology standards and intellectual property system: It will be implemented in two consecutive steps: (1) Work on interoperability security, traceability, and availability; (2) Set and improve AI related to the network security, interoperability, privacy protection, industry applications, and other technical standards. The directive aims for industry to participate on developing international standards, with the objective of fostering overseas applications.

- Establish an AI security supervision and evaluation system: AI security supervision should be conducted not only to assure national security and secrecy, but also to secure protection of people, technology, material, and management support. AI security monitoring will construct an early-warning mechanism, which together with an AI technology prediction will grasp the technology and industry trends. Restraint guidance and prospective prevention should be strengthened by a tailored risk assessment study, together with its prevention and control. China, like the EU, considers the implementation of accountability. It will be done by a transparent and open AI supervision system. A two-tiered regulatory structure will be applied to the supervision of the whole process. China claims to promote self-discipline. Evaluation mechanisms will be developed on AI, together with indicators system and systematic testing methods. Finally, the AI security certification will be done on a cross-domain AI test platform with systems key performance and assessment of AI products.

A decree of the president of the Russian Federation on the development of AI affirms that: "Creation of an Integrated System for Regulating the Social Relations Arising in Line with the Development and Use of Artificial Intelligence Technologies" [19]. It will be based on mainly on the following points applied to the technological solutions developed on the basis of artificial intelligence: ensuring conditions for access to data; ensuring favorable legal conditions for access to data; formulating ethical rules for human interaction with artificial intelligence; eliminating administrative barriers during the exportation of civilian products; creating legal conditions and establishing procedures for the simplified testing and introduction of technological solutions, as well as delegating the possibility of individual decision-making to information systems (with the exception of decisions that might infringe upon the rights and legitimate interests of individuals); creating unified systems for the standardization and assessment.

The legal conditions are planned to be implemented by 2024, and by 2030 a flexible legal regulatory system must be functioning.

Japan has created the "Strategic Council for AI Technology" [20]. lFive National Research and Development Agencies will be managed by the Council, under the jurisdiction of the Ministry of Internal Affairs and Communications, Ministry of Education, Culture, Sports, Science and Technology, and Ministry of Economy, Trade and Industry. Three research centers will depend on the following agencies: Center for Information and Neural Networks (CiNet) and Universal Communication Research Institute (UCRI) of the National Institute of Information and Communications Technology (NICT); RIKEN Center for Advanced Intelligence Project (AIP) of the Institute of Physical and Chemical Research (RIKEN); Artificial Intelligence Research Center (AIRC) of the National Institute of Advanced Industrial Science and Technology (AIST).

There are two institutions to implement the projects, the Japan Science and Technology Agency (JST) and New Energy and Industrial Technology Development Organization (NEDO). The strategy is based on nine principles while the strategy of the EU is based on seven [17]. According to the accountability, "Developers should make efforts to fulfill their accountability to stakeholders including AI systems' users."

In other words, the developers will provide explanations and information to users/providers about the AI system's characteristics with active involvement of stakeholders. There is a policy proposal about harmonization of AI and regulations, strategically taking the initiative in international standards and holding intellectual property.

The Institute of Electrical and Electronics Engineers (IEEE) Global Initiative on Ethics of Autonomous and Intelligent Systems can serve as a baseline to address requirements for accountability of Energy Industry. **The IEEE P7000** series is focused on ethical concerns in technological issues [8]. IEEE centers the accountability on: autonomous and intelligent technical systems; government and industry stakeholders; the manifestations generated by autonomous and intelligent technical systems. The accountability should be Ethically Aligned Designed (EAD) with universal human values, political self-determination data agency, and technical dependability. The chapters of EAD apply the general principles to the practice by a series of recommendations: General Principles; Affective Computing; Methods autonomous and intelligent systems (A/IS) design; A/IS for Sustainable Development; Embedding Values into A/IS; Policy; Law. According to [20]: "A/IS shall be created and operated to provide an unambiguous rationale for decisions made." Appendix II shows the standard projects that are under development.

2.3. Responsible Applications of AI in the Energy Industry

Any AI application in the energy sector should involve responsibilities. The responsibilities should be clearly defined by laws, standards, rules, etc., and from design to the use of the AI product or services, including also development, procurement, deployment, operation, and, finally, validation of effectiveness [21]. The responsibilities should consider the following issues (K7):

- The code/hardware track record employed by the AI product. It will also support the transparency record, including the records from inside and outside of the developers, e.g., some codes can be open source or created by third parties.
- All of the roles and responsibilities of every agent involved in the value chain. This is required in order to assign liabilities and the proportion of them according to the faults.

- Following on from this, all agents will need to have the information and training necessary to understand their responsibilities and the legal consequences of their work.
- All of the rules and standards must be created and applied according to the laws and regulations system where the AI product could have any consequence or effect.

It is also desirable that anybody involved in the value process of the AI product preserve documentation relative to procedures, certifications, decisions, etc., for the period that should be defined in any cases by the responsible agency, in case that it can be required for any authority.

Section 2.2. shows that several countries have begun to work on this issue, creating agencies in order to ensure compliance.
Accountability should be clearly linked to the other six key requirements, mainly effectiveness, competence, and transparency. They cannot be considered individually; instead, they must be treated as a whole to ensure ethical and legal compliance, and the technical safety and reliability of the AI products.

The recommendations given by [14] are also based on the seven key requirements, such as:

1. Creators of AI products should define clearly the outcomes of the product, e.g., accuracies, validation, etc., and also their responsibilities (K2).
2. All of the agents involved in the operation of the AI product should understand the responsibilities and the potential legal liability. As mentioned above, the legal environment it is not yet sufficiently clear for developers to guard against liability and governments are working on this to improve it (K3).
3. Responsibilities should be clearly defined in the contracts (K7).
4. Operators and creators should have the support of professional and independent mechanisms to guarantee the requirements of the AI products, e.g., agencies, audits, etc. (K1, K2)
5. The government, agencies, etc., responsible for legal conditions, standards, normalization, etc., should also create individually and collectively a variety of incentives to ensure that the outcomes of the AI product are in line with the ethical and accountability issues (K7).
6. Inquiries to determine responsibilities should consider the above points, considering all of the agents involved in the AI product and all the activities performed in the value chain of the product (K1, K7).

As can be seen in Fig. 3, *the most important applications of AI in the Energy sector* are energy efficiency in smart buildings, energy utilization/power consumption, energy storage, and smart grids.

AI becomes increasingly more important in the energy industry with a great potential for the future design of energy systems. AI can help the energy industry to become more efficient and secure, for instance by analyzing and evaluating the large volume of data.

AI in the power grid and smart grids (K1, K3): AI is present in the field of intelligent networking of electricity generation and consumption. With the increasing decentralization and digitalization of the power grid it is more difficult to manage the grid participants keeping the grid in balance in the same time. AI can be used here for evaluating and analyzing the big amount of data as quickly and efficiently as possible. In smart grids, where power plant generation has dramatically increased, AI can help to evaluate, analyze, and control the data of different actors, such as consumers, producers, and energy storage systems. A particular case for smart grid applications is integrating EVs. AI can help to monitor and coordinate the charging of EVs, thereby offering the possibility of storing electricity and stabilizing the grid.

AI applications for variable RES integration (K2, K3, K6): The potential of AI is being unlocked by the generation of big data and increased processing power enabling fast and intelligent decision making. This leads to increased grid stability and reliability as well as flexibility by integration of variable RES, like wind and solar, using for instance optimized energy storage solutions. Digitalization and digital technologies can better support RES integration in smart grids.

AI in electricity trading (K2): AI in power trading can help to improve forecasts by systematically evaluating large amounts of data, such as weather data and historical data. Better forecasts can also increase grid stability and, thus, security of supply. AI can also help to automatically monitor and analyze trading on the electricity market thereby enabling the detection of faults preventing deviations from standard values.

AI used for improving the energy efficiency in smart buildings (K3): AI can monitor and control power consumption using smart solutions based on smart sensors and smart meters contributing to a more stable and green electricity grid. Analyzing data-based user preferences leading to informed responses in the electricity market could save electricity, thus reducing the cost.

AI applications for energy accessibility (K4): Smart home applications-based AI, such as Verv and PowerScout [8], can assist users with energy management. These kinds of applications enable users to monitor records on how each appliance in their homes uses energy and to regulate their energy expenses. These applications also have safety features and provide tips for reducing carbon emissions, thereby helping clients in making the right decisions when deploying RES for their homes.

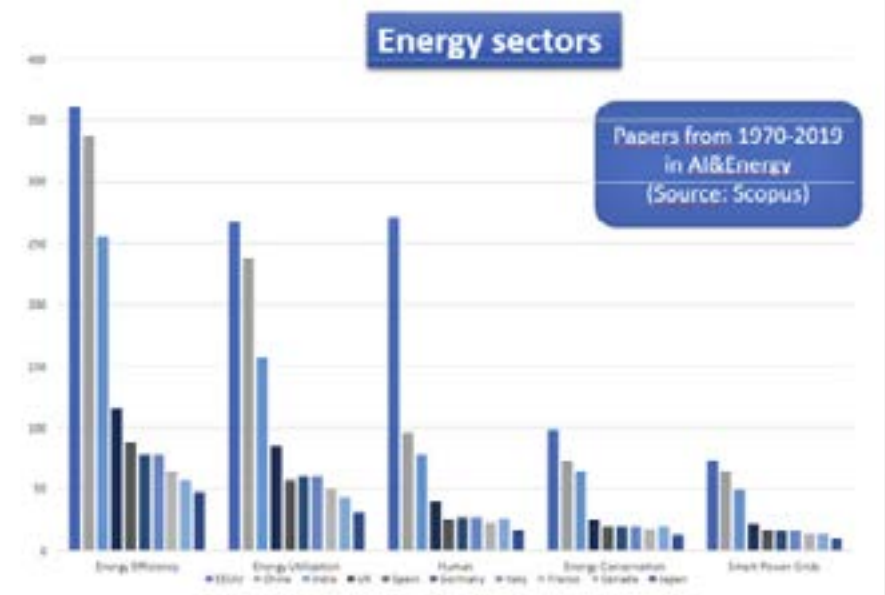


Fig. 3. The most important applications of AI in energy sector based on SCOPUS publications.

2.4. Contribution to Energy Sector Transformation and Challenges

The energy sector is undergoing a major transformation with the increase of RES technology (solar and wind) that provides variable energy, distributed energy resources (DERs), bidirectional power flow of electricity, large flows of data collected by IoT and other devices, increased use of energy storage and the evolving role of utilities and consumers.

According to the EC's white paper on AI [4], the competent authorities should be in a position not only to investigate individual cases, but also to assess the impact on society. The trustworthiness and security of AI, based on European rules and values, must be effectively enforced by affected parties and the European and competent national authorities. Procedures for certification, inspection, or testing, including prior conformity assessment, must be carried out because of the high risk of some AI applications. For example, checking the datasets used in the development phase and the algorithms, in the conformity assessment mechanisms that already exist.

In cases where these mechanisms do not exist, similar mechanisms may need to be established, considering the inputs of the European standards organizations, stakeholders and the best practices. The AI must be non-discriminatory and proportionate, employing objective and transparent criteria in accordance with international obligations. It would be mandatory for all economic operators addressed by the requirements the conformity assessments.

The EC establishes some support structure in order to limit the burden on SMEs, through the Digital Innovation Hubs, dedicated online tools, and standards. According to the white paper "Any prior conformity assessment should be without prejudice to monitoring compliance and ex post enforcement by competent national authorities." Competent authorities and third parties can test AI applications by ex-post controls. Therefore, they should be enabled by correct documentation. A continuous market surveillance scheme will be required for the compliance monitoring. Finally, effective judicial redress should be ensured for parties who are negatively affected by AI systems.

2.5. Barriers and Risks to AI Adoption

There are not yet clear rules, laws, or standards, etc. that the agents involved can apply. Therefore, it is difficult or not possible to assign any responsibility regarding AI products or services. According to [15], if a product is not assigned any responsibility, then it cannot be trusted.

AI is a product that can be offered as a "black box," i.e., opaque, generating a high risk to the agents involved. For example, it can have complex algorithms with a large number of codes, and the algorithms can change over the time to adapt to the data inputs or the outcomes. The data can have a large number of variables, volume, variety, complexity, etc. The process is also complex, with many agents involved in the process, e.g., programmers, engineers, data analysts, owners, operators, etc. involved in the value chain, from the design to the end of the life cycle of the AI product. Thus, it generates a great challenge in order to set responsibilities and accountability.

It has been shown in Section 2.1 that countries worldwide are working on this issue, to create agencies to control AI, develop standards, etc. to guarantee accountability, but it is not done yet. It must set clear responsibilities to verify agents along the value chain of the AI product.

Data protection and data security are some of the most prominent weak points of the use of AI. In 2018, the German Federal Office for Security (BSI) observed that the number of cyberattacks on critical infrastructure tripled in comparison with the previous year. This is why cybersecurity is becoming increasingly important in order to protect the power grid [7].

Many end users are critical to AI, especially in relation to smart home technologies. The biggest obstacle to the acceptance of smart meters is fear of revealing private information without knowing exactly how it is used. There is still no regulation in many countries on how to handle these sensitive data.

Chapter 3.
What the Energy Sector Must Do to be Compliant with the Seven Key Requirements

According to [21], the cost reduction in the renewable energy industry, mainly based on wind turbines and solar photovoltaics, is given by the Paris Agreement [23] and the energy scenarios [22]. The European Union has set an objective to reduce the total domestic greenhouse gas (GHG) emissions by 80% in 2050 compared to 1990 levels.

3.1. Concrete and Practical Steps that the Energy Sector Must Take to be Compliant

AI systems and their outcomes should be defined and detailed in the contracts, i.e., an agreement between private parties creating mutual obligations enforceable by law. Law includes also "private law," i.e., the terms of the agreement between the parties who are exchanging promises. Therefore, it should, meet the EU's safety rules (existing as well as possible complementary ones). There are no standards yet, but they are under development by, for example, ISO and IEEE (see Appendices). The contract/agreement should consider almost all of the following issues that the law will enforce:

- Bodies: almost all of the property and contractor
- Insurance: the contractor must contract insurance for safety, and insurance(s) must include to the property.
- Responsibilities: responsibilities are all from the contractor and the contractor shows to the property the organizational chart and the responsibilities.
- Standards and laws
- Guaranties and penalties
- Must be fixed in the appendix of the contract. They are applied to the contractor
- The contractor must report enough detailed information to the property weekly and monthly by detailed report about the AI activities, dataset and any incidents
- Any new AI activity by the contractor must be reported to the property and to have its approbation.
- Any loss, theft, accident etc. regarding to AI hardware/software, i.e., AI system, is the responsibility of the contractor.
- It should be desirable to have the support of professional and independent mechanisms to guarantee the requirements of the AI products, e.g., agencies, auditors, etc.

According to the EU [13], the contract should also include (see also Section 2.2):

--The use of the products or services where the AI system is applied must be safe.
--The datasets should be available upon request for any competent authorities, e.g., inspections.
--The process should protect confidential information (e.g., trade secrets).
--The datasets employed to train the AI systems need to consider all of the scenarios required and to be sufficiently broad.
According to the IEEE, the following should also be considered [24]:

- The code/hardware track record employed in the AI product.
- Definitions of the roles and responsibilities of every agent involved in the value chain.
- According to the previous point, all agents will need to have the information and formation needed to understand their responsibilities and the legal consequences of their work.
- All of the rules and standards must be created and applied according to the laws and regulations system where the AI product could have any consequence or effect.

It is also desirable that everybody involved in the value process of the AI product preserve documentation related to procedures, certifications, decisions, etc., almost for the period that should be defined in any case by the responsible agency, in case that it can be required for any authority.

Concrete and practical steps for AI Technology in improving RES based on the seven key requirements:

--**Smart-grid-based energy storage optimization solutions (K1)**; Energy storage systems, in terms of large-scale, aggregated small home battery or plugged-in electric vehicles are key solutions for renewable integration. AI can support these solutions more efficiently, thereby maximizing RES integration, including the reduction of forecast errors, minimizing prices for electricity consumers, and maximizing returns for the owners of the storage systems. The speed of complexity of managing energy storage systems in a dynamic environment requires advanced AI techniques and algorithms, which are able to extend the battery life.

--**Improving the integration of (hybrid) microgrids (K2)**: AI can help with the integration of microgrids and managing distributed generation. The AI-powered control system can play a vital role in solving the quality and congestion issues, balancing the energy flow within the hybrid microgrids (AC and DC).

--**Better algorithms for energy forecasting (K1, K2)** for PV systems and wind turbines integrated into power systems. Big data, machine learning (ML), and AI can produce accurate power generation forecasts that will make it feasible to integrate

much more renewable energy into the grid. For system operators, accurate forecasting can improve unit commitment, thereby increasing dispatch efficiency and reducing reliability issues.

--**Digitalization to support the energy sector** (based on the European Green Deal): AI and ML become increasingly more important in the energy sector making the industry more efficient and secure by analyzing and evaluating the large amount of data faster and more accurately. Typical areas of application are electricity trading, smart grids, heating, and transport, etc. Digital technologies can support the energy sector in several ways, including better monitoring, operation, and maintenance of RES assets, more system operation and real-time control strategies, and implementation of new market design etc.

--**Integration of electromobility- K2, K4** AI and big data techniques can ease the full integration of the electrified mobility with the smart grid and RES by enabling an opportunistic charging strategy where state-of-charge of on-board battery energy storage (BES) is monitored and charging/discharging phases are coordinated. This way the smart grid does not collapse (avoiding excessive simultaneous requests of vehicle recharging), vehicle recharging time can be minimized, and there is the possibility (e.g., in parking areas) of bidirectional energy transfer, i.e., from one vehicle BES to another, or from vehicle BES to the grid.

--**Coordination of maintenance work and determination of optimal times** for maintenance of network (K2). This helps in minimizing costs and losses. For example, by detecting anomalies in generation, consumption, or transmission, AI can stabilize the power grid by developing suitable solutions.

--**Software platforms and tools (K1)** that leverages AI to optimize energy consumption: OPTIMAX from ABB [25] is a scalable and highly flexible energy management industrial platform, which can easily be integrated into existing and complex infrastructures to improve energy efficiency in smart buildings, smart transportation, and smart homes.

--**The introduction of advanced and intelligent technologies into the power sector needs to be weighed up against cybersecurity. Policymakers need to strike a balance between supporting the development of AI technologies and managing any risks from malicious actors (K7).** AI algorithms and tools can improve their knowledge so as to understand threats and cyber risks, minimizing and reporting the negative impacts.

3.2. Case Studies in Energy Sector Considering the Ethical Principles and Guidelines

Building an in-house infrastructure from the ground up is costly, heavy on resources, and largely dependent on specific skills many industrial companies lack. But there is a solution: industrial IoT platforms that give a solid foundation and necessary tools to

unroll predictive maintenance activities. Varying in scope and the set of features offered, they usually provide companies the following capabilities and services: device management to connect hundreds of thousands of sensors and meters on one platform; support for industrial messaging protocols; software development environment, tools, and APIs to integrate with existing enterprise solutions; scalable data storage and a big data processing engine; analytics engines and machine learning as a service; Digital Twin technology — visualizations of the equipment's condition in real-time; and ready-to-use asset management software and analytical engines tailored for tasks.

Case studies:

--**Smart-grid-based energy storage optimization solutions (K1)**. For example, in Australia Tesla's battery in its first year of operation generated an estimated USD 24 million in revenue, while also providing a reduction of between USD 40-50 million in frequency control ancillary service costs [26].

--**Better algorithms for energy forecasting (K1, K2)**. IBM was able to show an improvement of 30% in solar forecasting while working with the U.S. Department of Energy's Sun Shot initiative in 2015. A successful example for accurate variable RES forecasts is that of the EWeLiNE research project using ML-based software in Germany in 2017 [27].

--**Predictive maintenance (K2)** based on AI and ML technologies for: early fault detection by using real-time platforms (for using wind farm and solar plant data), equipment failure, condition monitoring systems and production reliability. The goal of predictive maintenance is to optimize the balance between corrective and preventative maintenance, by enabling just in time replacement of components. This approach replaces only those components when they are close to failure. By extending component lifespans (compared to preventive maintenance) and reducing unscheduled maintenance and labor costs (over corrective maintenance), businesses can gain cost savings and competitive advantages.

An example of predictive maintenance applied in the energy industry: predictive analytics can take sensor data from a wind turbine or PV system to monitor the components and to predict with high accuracy when the system need maintenance. GE in Japan with the help of AI succeeded in enhancing wind turbine efficiency, reducing maintenance costs by 20% and increasing power output by 5% [4].

--Energy consumption profiling: The issue of big energy companies using AI and big data to track consumers' behavior is critical: on one side there is the opportunity for an accurate (in space and time) estimation of energy demand to optimize the trade-off among energy comfort for the users, energy efficiency for the grid, and energy saving for the planet. On the other side, there is the risk of unbalanced power between

the two parties, big energy company vs. EU citizen as an energy user, with threats of privacy and confidentiality violation for the latter (K1, K2, K6). Hence, regulation related to big data and AI use for energy consumption profiling should ensure "equal power" to data access and management.

--Local energy communities have the right to access and manage the energy they created (*K*1, *K*4, *K*6). Concerned by the events at Chernobyl in 1986, the citizens of Schönau ([28]), a village of approximately 4,400 inhabitants in the German state of Baden-Württemberg, proposed more reliance on RES instead of nuclear power plants. When their idea was dismissed by the local energy provider, the citizens sought to take back the franchise for power supply [29]. They had to do so over the objections of their own local government, which had obtained a franchise fee from the provider. After two referendums, the citizens founded a cooperative called Elektrizitätswerke Schönau (EWS) in 1996 that was granted the concession. Instead of relying on coal and nuclear power as the incumbent utility had done, EWS's members learned how to use solar PV and geothermal power to supply Schönau. Facilitated by the liberalization of the electricity industry around this time, EWS also began to win concessions in other parts of the country and currently operates nine electricity and gas networks in Germany. In addition to its own generation, EWS also purchases renewable energy from independent producers in Austria, Germany, and Scandinavia to supply its end-use consumers all while maintaining a cooperative governance structure. This is related to K4 from Chapter 2 regarding transparency: EWS not only provides its customers with a certification of how their energy is sourced but also involves its members in its decision-making processes. While the desire for a more environmentally friendly energy system instigated the formation of EWS, its success ultimately hinged on regulation that was crafted to break the power of incumbent monopolies and provided a level playing field. Nevertheless, EWS may never have been launched if its founders had been unable to buy out the concession from the incumbent at the appraised value of DM 8.7 million even though another appraiser had it valued at DM 4 million. This is related to K6 from Chapter 2 about reducing barriers to entry for environmental solutions: as a pioneer, EWS had to litigate its way into existence and was fortunate that its takeover of the Schönau franchise coincided with electricity industry liberalization in 1998.

Thus, it was able to become financially viable by expanding to the extent that it now supplies some 170,000 consumers [28]. Future environmental solutions should not face such barriers, and regulation for them could be based on the existing Article 16 about energy communities in the EU Directive 2019/944 [13] on the internal market for electricity. In particular, it prevents Member States from barring local energy communities from access to energy markets and ensures that their financial settlements take place in a transparent manner. Moreover, the implementation of a trustworthy AI paradigm can be an opportunity to facilitate the creation of a community of prosumers, i.e., local energy communities where citizens act as both energy producers and consumers and have the right to access and manage the energy they created.

--**A network platform – K1, K6** (IoT middleware platform-based cloud computing) or infrastructure to be able to exchange the energy produced. Verdigris Technologies offers a software platform that leverages AI to optimize energy consumption in commercial buildings. Some hotels in San Francisco, USA, that used the app to identify energy insufficiency in their commercial kitchens confirmed that within three months of using the application inefficiencies were identified that were costing them over USD 13.000 in preventable annual losses.

--**Data protection by default and data protection by design (K3)**. Some social issues, such as diversity, nondiscrimination, and fairness, need to be taken into account in the technical design process and implementation of AI. Article 25 of the GDPR requires organizations to implement both data protection by design and by default. This means that legal and compliance issues need to be considered from the beginning of a project, and that privacy must be a deciding factor in implementation choices.

However, detailed investigation of smart grid projects reveals that practice sometimes falls short of this. For example, Brown [30] examined the British smart meter program and found that little attention was paid to privacy in the early phases of its development. It became a more significant feature of the design after a public consultation process. Similarly, Murphy [31] found that the importance of privacy by design was not fully appreciated in the early stages of the Irish smart metering program, and it was neither a key design principle nor an evaluation criterion. Privacy was seen as something to be balanced against other competing interests, in a zero-sum fashion, which is a contradiction of the privacy by design concept. However, Cavoukian [32] highlights how some providers are engaging fully with privacy, such as the German supplier Vattenfall, which included a data privacy representative throughout all stages of the design of their smart meters, and has highly restricted the use of data from the system; and the grid operator Alliander, which invested heavily to obtain a Data Privacy and Security certification as part of a consumer reassurance campaign.

--**IBM, Microsoft, and the AEE Institute** in the last two years tried to respond to **emerging threats developing different hardware and software solutions against cyberattacks**, which can be used to automate and secure demand-side management based on IoT, thereby minimizing the negative impacts (K7). Protections under development aim to make the distribution electricity system more resilient against cyberattacks. This meets the EU's safety rules (existing as well as possible complementary ones) and the standards set. These solutions aim to guarantee the AI with regard to network security, interoperability, privacy protection, industry applications, and other technical standards. In the case of China, it is in accordance with its plan to promote the self-discipline in the industry. These initiatives aim to reduce the damaged caused by AI systems. Some other macro-initiatives are being carried out in the EU [10], e.g., France has established a group of experts to verify algorithms and databases and to improve understanding in civil society, thereby creating a consultative ethics committee

for digital technologies and AI, which would organize public debate in this field. In a similar vein, the U.K. has created a Centre for Data Ethics and Innovation that will be tasked with ensuring safe, ethical, and ground-breaking innovation in AI and data-driven technologies, etc.

Chapter 4.
Conclusion, Practical Recommendation and Obligations

Practical recommendation and obligations emerge from the analysis on how to utilize the laws, standards, and regulation to accommodate the new capacities, practices, and behaviors. Many of these recommendations and obligations are underpinned by the GDPR, which requires respect for fundamental rights (recital 2), adequate security (Article 5 (1)(f)), good data governance (Article 26), transparency (Articles 12 to 14), fairness (Article 5(1)(a)), and accountability (Article 5(2)).

Practical recommendation for Energy Industry based on the seven key requirements (K1-K7):

K1. Recommendation for **Human agency and oversight – including fundamental rights, human agency, and human oversight:**

> Uses of AI in the energy industry must take account of the fact that individuals and groups will have divergent but still legitimate goals, and allow the possibility of exercising choice. AI must enable better, more informed human decision-making, not take away control.
> RES must be designed so that unusual events or fluctuations are brought to the attention of operators as quickly as possible, and intervention is always possible, so that runaway algorithms or cascading failures are avoided. Human oversight is also an important mechanism for identifying security breaches and risks, and therefore systems should be both auditable and regularly audited.

K2. Recommendation for **Technical robustness and safety – including resilience to attack and security, fall-back plan and general safety, accuracy, reliability, and reproducibility:**

> Predictions based on AI techniques can be used not only for energy production, storage and distribution, but also for energy trading, e.g., to find the right compromise between energy production-selling price-user consumption in a free but transparent energy market. AI should contribute to the improvement of safety and technical robustness, as well as the identification and mitigation of attacks.

Fingerprinting and anomaly detection is the first step to classify faults and cyber-attacks correctly thereby providing a fall-back plan to ensure safety of the energy system.
AI-based fingerprinting and anomaly detection can be used for not only energy production, storage, and distribution plants, but also to detect anomalies during energy trading sessions for a transparent energy market.

K3. Recommendation for **Privacy and data governance – including respect for privacy, quality, and integrity of data, and access to data:**

To give the energy industry and in particular end-use consumers more confidence in AI (e.g., in relation to smart home technologies-smart meters or optimization of charging infrastructure in smart recharging grid for electrified mobility), it must be clearly communicated how the data are used and by whom, while data security must be guaranteed.
In addition to compliance with ethical guidelines, it is essential that AI in the energy sector complies fully with Standards (IEEE P7002) and European law, particularly the General Data Protection Regulation (GDPR), and applies the Data Protection Impact Assessment Template prepared by the Commission's Smart Grid Task Force. There should be clear security requirements for participants in the electricity market to ensure data confidentiality, integrity, and authentication along the whole chain of data acquisition, processing, and storage. This can be achieved by defining best practices to be followed for security of data taking note of the different roles: users, production industries, companies working as service providers and institutions (standardization bodies, regulatory authorities, and governments).

K4. Recommendation for **Transparency – including traceability, explainability, and communication:**

Marshaling this existing expertise together with algorithmic advances in inverse optimization could be used by auditors to unveil how AI applications function in the energy sector and to ensure that their use is in line with social objectives.

K5. Recommendation for **Diversity, non-discrimination and fairness – including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation:**

In order to ensure non-discriminatory programming and functioning, systems need to be trained and tested for unfair bias. Therefore, companies should test their algorithms for bias and discrimination and demonstrate that certain fairness standards are met. The IEEE P7003 standard can serve as a baseline to address and eliminate issues of harmful bias in development of AI algorithms.

K6. Recommendation for **Societal and environmental wellbeing – including sustainability and environmental friendliness, social impact, society, and democracy:**

AI in power trading helps to improve forecasts (based on weather data, historical data, consumption profiles and mobility patterns) as well as facilitate and speed up the integration of RES (green energy). Better forecasts also increase grid stability and thus security of supply.
Regulatory authorities' use of AI would add another layer of intelligence that could anticipate the actions of market participants and tweak market designs to discourage deviations from societal objectives.

K7. Recommendation for **Accountability – including auditability, minimization, and reporting of negative impact, trade-offs, and redress:**

Accountability should be clearly linked to the other six key requirements, mainly effectiveness, competence, and transparency. They cannot be considered only individually; instead, they must be treated as a whole to ensuring ethical and legal compliance, and the technical safety and reliability of the AI products.
The Global Initiative on Ethics of Autonomous and Intelligent Systems from the Institute of Electrical and Electronics Engineers **(IEEE)** can serve as a baseline to address requirements for accountability of the energy industry. For instance, the **IEEE-P7001** standard can serve as a baseline to address requirements for transparency and accountability of the energy sector.
Energy sector companies should transparently communicate and report negative impacts of the AI products.

References

[1] European Commission, "Building trust in human-centric artificial intelligence" (Communication from the Commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions) COM (2019) 168.es

[2] European Commission, "Artificial intelligence for Europe" (Communication from the Commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions) COM (2018)_237e.n

[3] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy Artificial Intelligence" 2019. [Online]. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation

[4] European Commission, "On Artificial Intelligence – A European approach to excellence and trust" 2020 t. [Online]. Available: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[5] Smart Grid Task Force Expert Group 2, "Data Protection Impact Assessment Template for Smart

Grid and Smart Metering systems" (2018). [Online]. Available: https://ec.europa.eu/energy/sites/ener/files/documents/dpia_for_publication_2018.pdf

[6] Smart Grids Task Force Expert Group 1, "My Energy Data" (2016). [Online]. Available: https://ec.europa.eu/energy/sites/ener/files/documents/report_final_eg1_my_energy_data_15_november_2016.pdf

[7] Smart Grid Task Force Expert Group 2, "Recommendations to the European Commission for the Implementation of Sector-Specific Rules for Cybersecurity Aspects of Cross-Border Electricity Flows, on Common Minimum Requirements, Planning, Monitoring, Reporting and Crisis Management" (2019). [Online]. Available: https://ec.europa.eu/energy/sites/ener/files/sgtf_eg2_report_final_report_2019.pdf

[8] Institute of Electrical and Electronics Engineers, "Ethics in Action in Autonomous and Intelligent Systems" https://ethicsinaction.ieee.org (accessed Sep. 17 2020).

[9] R. Fernández-Blanco, J. M. Morales, S. Pineda, Á. Porras, "Kernel-Based Inverse Optimization: Application to the Power Forecasting and Bidding of a Fleet of Electric Vehicles". arxiv.org. https://arxiv.org/abs/1908.00399 (accessed Sep. 17 2020).

[10] S. Samuel, "A new study finds a potential risk with self-driving cars: failure to detect dark-skinned pedestrians". Vox.com. https://www.vox.com/future-perfect/2019/3/5/18251924/self-driving-car-racial-bias-study-autonomous-vehicle-dark-skin (accessed Sep. 17 2020).

[11] M.A. Van Sluisveld, A.F. Hof, S. Carrara, F.W. Geels, M. Nilsson, K. Rogge, B. Turnheim, and D.P. van Vuuren, "Aligning integrated assessment modelling with socio-technical transition insights: An application to low-carbon energy scenario analysis in Europe," Technological Forecasting and Social Change vol. 151, pp. 119177, 2020.

[12] Nasdaq, "Nasdaq Renewable Index Wind Germany". nasdaqomx.com. http://www.nasdaqomx.com/transactions/markets/commodities/renewables (accessed Sep. 17 2020).

[13] Directive (EU) 2019/944 of the European Parliament and of the Council of 5 June 2019 on common rules for the internal market for electricity and amending Directive 2012/27/EU. OJ L 158, pp. 125–199, 14 Jun. 2019.

[14] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC. OJ L 199, pp. 1–88, 4 May 2016.

[14] L. Delponte and G. Tamburrini, "European artificial intelligence (AI) leadership, the path for an integrated vision," European Parliament, 2018.

[15] Directive (EC) 2001/95 of the European Parliament and of the Council of 3 December 2001 on general product safety. OJ L 11, pp. 4–17, 15 January 2002.

[16] Executive Order 13859 on maintaining American leadership in artificial intelligence. 2019.

[17] NIST. "U.S. Leadership in AI: A plan for federal engagement in developing technical standards and related tools", National Institute of Standards and Technology. U.S. Department of Commerce, 2019.

[18] G. Webster, R. Creemers,; P. Triolo, and E. Kania, "Full translation: China's 'new generation artificial intelligence development plan' (2017)". Digi China, August 2017, 1.

[19] Russian Federation, Decree of the President of the Russian federation on the development of artificial intelligence in the Russian Federation. Office of the President of the Russian Federation 2019.

[20] The Conference toward AI Network Society. Draft AI R&D guidelines for international discussions. The Conference toward AI Network Society 2017.

[21] B. Steffen, M. Beuse, P. Tautorat, and T.S.Schmidt, “Experience curves for operations and maintenance costs of renewable energy technologies,” Joule, vol. 4, pp. 359–375, 2020.

0ys[22] IRENA 2019-International Renewable Energy Agency, Innovation landscape brief: Artificial Intelligence and big data, Abu Dhabi, Report available online: www.irena.org/publications

s.g[23] J. Rogelj, M. Den Elzen, N. Höhne, T.Fransen, H. Fekete, H. Winkler, R. Schaeffer, F. Sha, K. Riahi, and M. Meinshausen, “Paris agreement climate proposals need a boost to keep warming well below 2∞C,” Nature vol. 534, pp. 631-639, Jun. 20169.

[24] IEEE Standards Association, “The IEEE global initiative on ethics of autonomous and intelligent systems.” IEEE.org. https://standards.IEEE.Org (accessed Sep. 23, 2020).

[25] Optimax - ABB Mission to Zero, www.abb.com/mission-to-zero-optimax

[26] M. Mezengarb (2019), “Tesla big battery paves way for artificial intelligence to dominate energy trades.” Renew Economy, http://reneweconomy.com.au/tesla-big-battery-paves-way-for-artificial-intelligence-to-dominate-energy-trades-31949 (accessed Sep. 23, 2020).

[27] EWeLiNE research project, “Renewable energy generation forecasting”, Germany.

[28] https://www.ews-schoenau.de/oekostrom/

[29] https://www.next-kraftwerke.com/knowledge/artificial-intelligence. What is Artificial Intelligence in the Energy Industry?

[30] I. Brown, “Britain's smart meter programme: A case study in privacy by design,” International Review of Law, Computers & Technology vol. 28, pp. 172–184, 2014.

[31] M. H. Murphy, “The Introduction of Smart Meters in Ireland: Privacy Implications and the Role of Privacy by Design,” Dublin University Law Journal vol. 38, pp. 191–207, 2015.

s[32] A. Cavoukian, “Smart meters in Europe: Privacy by Design at its best,” Information and Privacy Commissioner of Ontario, 2012.

. . .. n. . . 50H. e 1. s) xI n 44 6,l.6A t ann5l g 2

Appendix I. Standard and/or project under the direct responsibility of ISO/IEC secretariat related to AI

•ISO/IEC CD TR 20547-1: Information technology — Big data reference architecture — Part 1: Framework and application process

•ISO/IEC CD 22989: Artificial intelligence — Concepts and terminology

•ISO/IEC CD 23053: Framework for AI Systems Using Machine Learning (ML)

•ISO/IEC AWI 23894: Information Technology — Artificial Intelligence — Risk Management

•ISO/IEC AWI TR 24027: Information technology — Artificial Intelligence — Bias in AI systems and AI aided decision making

•ISO/IEC PRF TR 24028: Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

•ISO/IEC CD TR 24029-1: Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview

•ISO/IEC CD TR 24030: Information technology — Artificial Intelligence (AI) — Use cases

•ISO/IEC AWI TR 24368: Information technology — Artificial intelligence — Overview of ethical and societal concerns

•ISO/IEC AWI TR 24372: Information technology — Artificial intelligence (AI) — Overview of

computational approaches for AI systems

•ISO/IEC AWI 24668: Information technology — Artificial intelligence —Process management framework for Big data analytics

•ISO/IEC AWI 38507: Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations

Appendix II. IEEE P7000 standards projects

The following standard projects are being done:

• IEEE P7000 - IEEE Standards Project Model Process for Addressing Ethical Concerns During System Design.

• IEEE P7001 - IEEE Standards Project for Transparency of Autonomous Systems.

• IEEE P7002 - IEEE Standards Project for Data Privacy Process.

• IEEE P7003 - IEEE Standards Project for Algorithmic Bias Considerations.

• IEEE P7004 - IEEE Standards Project for Child and Student Data Governance.

• IEEE P7005 - IEEE Standards Project for Employer Data Governance.

• IEEE P7006 - IEEE Standards Project for Personal Data AI Agent Working Group.

• IEEE P7007 - IEEE Standards Project for Ontological Standard for Ethically Driven Robotics and Automation Systems.

• IEEE P7008 - IEEE Standards Project for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems.

• IEEE P7009 - IEEE Standards Project for Fail-Safe Design of Autonomous and Semi-Autonomous Systems.

• IEEE P7010 - IEEE Standards Project for Well-being Metric for Autonomous and Intelligent Systems

• IEEE P7011 - IEEE Standards Project for the Process of Identifying and Rating the Trustworthiness of News Sources

• IEEE P7012 - IEEE Standards Project for Machine Readable Personal Privacy Terms

• IEEE P7013 - IEEE Standards Project for Inclusion and Application Standards for Automated Facial Analysis Technology

4

HEALTHCARE

Applying a Human-Centered Approach to Assess Risks of Using AI Systems in Healthcare

Authors

Raja Chatila
Chairman Healthcare Committee, AI4People; Professor and Director of the Institute of Intelligent Systems and Robotics (ISIR) at Pierre and Marie Curie University in Paris, France

Stephen Cory Robinson
Senior Lecturer/Assistant Professor in Communication Design at Linköping University, Norrkoping, Sweden

Donald Combs
Vice President & Dean of the School of Health Professions, Eastern Virginia Medical School, USA

Paula Boddington
Senior Research Fellow, New College of the Humanities London, UK

Hervé Chneiweiss
Directeur de Recherche au CNRS, Paris, France

Eugenio Guglielmelli
Senior Advisor on Publications for IEEE RAS Professor of Bioengineering Prorector for Research Founder, Research Unit of Biomedical Robotics and Biomicrosystems Università Campus Bio-Medico di Roma

Danny van Roijen
Digital Health Director at COCIR

Jos Dumortier
Honorary Professor of ICT Law at the University of Leuven, Belgium

Leonardo Calini
Policy Manager, European Government Affairs at Microsoft

1. Introduction

On April 8, 2019, the High-Level Expert Group on AI (HLEG-AI) appointed by the European Commission issued the "Ethics Guidelines for Trustworthy AI" (European Commission, 2019a). On June 26, 2019 the group issued the "Policy and Investment Recommendations for Trustworthy AI" (European Commission, 2019b).

In its Ethics Guidelines, the HLEG-AI has identified seven requirements considered key for the design, development, deployment and use of AI systems. AI-based systems complying with these requirements would be considered to be trustworthy and aligned with a human-centered approach.

The HLEG advocated that these requirements become a necessary condition for the adoption of AI systems in Europe.

In its Whitepaper entitled "*On Artificial Intelligence - A European approach to excellence and trust*" released on February 19, 2020, the European Commission summarized its planned AI policy as including "*Policy options for a future EU regulatory framework that would determine the types of legal requirements that would apply to relevant actors, with a particular focus on high-risk applications*." (European Commission, 2020a)

The risk-based approach defined in the Whitepaper is actually based on a two-tier definition of risk.

To be considered "high-risk", the AI system must be deployed in a sector known to be high-risk, e.g., the healthcare sector. Second the AI system must be used within this sector in an application, which is itself considered high-risk.

A subsequent report by the HLEG-AI, "*The Assessment List for Trustworthy Artificial Intelligence*" (ALTAI), was published on July 17, 2020 in an effort to provide an initial, more concrete, approach to evaluating compliance with the Ethics Guidelines for Trustworthy AI systems (European Commission, 2020b). The HLEG-AI stated that:

> "*The Assessment List for Trustworthy AI (ALTAI) is intended for flexible use: organisations can draw on elements relevant to the particular AI system from this Assessment List for Trustworthy AI (ALTAI) or add elements to it as they see fit, taking into consideration the sector they operate in. It helps organisations understand what Trustworthy AI is, in particular what risks an AI system might generate, and how to minimise those risks while maximising the benefit of AI. It is intended to help organisations identify how proposed AI systems might generate risks, and to identify whether and what kind of active measures may need to be taken to avoid and minimise those risks. Organisations will derive the most value from this Assessment List (ALTAI) by active engagement with the questions it raises, which are aimed at*

encouraging thoughtful reflection to provoke appropriate action and nurture an organisational culture committed to developing and maintaining Trustworthy AI systems. It raises awareness of the potential impact of AI on society, the environment, consumers, workers and citizens (in particular children and people belonging to marginalised groups). It encourages the involvement of all relevant stakeholders. It helps to gain insight on whether meaningful and appropriate solutions or processes to accomplish adherence to the seven requirements are already in place or need to be put in place."

The combination of the Guidelines and the Assessment List provide a solid foundation for the assessment of high-risk AI systems operating in high-risk sectors such as healthcare.

Following its past work on ethics and governance for AI (Floridi, 2018; Pagallo, 2019), AI4People has identified healthcare as one of the strategic sectors for the deployment of AI, and has appointed a working group to analyze how trustworthy AI can be implemented in this sector, which is considered high-risk. This paper examines how the seven requirements are relevant and can be used, along with other tools such as ALTAI, to assess risk in the deployment of AI systems in healthcare.

The aim is to illustrate a practical approach to assessing risk and to provide recommendations to stakeholders in this sector. It is important to note here that risk in healthcare, and thus in using AI in healthcare, is multi-dimensional. This multi-dimensionality will be explored through two examples of use cases that illustrate how risk can be assessed.

2. AI and healthcare

Healthcare is complicated. It involves assessing the current and trending state of health among patients with differing genetic makeups, personal history, environmental exposure, behavioral patterns, social contexts, cultures, economic status, self-awareness and patterns of healthcare usage.

Healthcare practitioners use various kinds of data for decision-making, including diagnostic laboratory and radiologic tests, written notes and electronic health records recounting patient interviews and anamnesis, data about the history of family health conditions, epidemiologic models of infectious diseases, and knowledge of available resources. They often use these data in situations of high urgency. The amount of data available from all these sources overwhelms the processing capacity of practitioners. It is therefore no surprise that AI systems find a great number of applications in this domain, where they promise faster and more comprehensive decision-making support

than practitioners can muster on their own. One substantial clinical application of AI has been in the imaging professions (Ting, 2018) - radiology and sonography - where, thanks to data availability and to progress in the development of effective algorithms, AI systems have shown a high level of accuracy, helping to identify tumors in breast cancer, retinal disease and recently even fast diagnosis of COVID-19 pulmonary diseases (McCall, 2020).

Recognizing the progress that has been made does not mean, however, that the AI systems should be already considered fully trustworthy. Generally speaking, many algorithms based on deep learning techniques are considered blalck boxes (Castelvecchi, 2016; Barredo Arrieta, 2020). Regarding interactions of practitioners and patients with AI systems, there is little understanding of the degree to which practitioners defer to the algorithms. And, finally, the big question of who is ultimately responsible for the diagnosis, treatment and outcomes of healthcare has not been answered yet.

3. Risk, danger and hazard

The notion of **risk** used in the EU Whitepaper (European Commission, 2020a) and in the ALTAI (European Commission, 2020b) needs to be clarified in order to analyze how to appropriately assess AI-based systems in healthcare.

A **risk** is a possible harm, more or less foreseeable, measurable by a probability of seeing a danger materialize, while the hazard is an unpredictable and unexpected event, even if it could be probabilistically modelled.

A **danger** is the presence of a factor that compromises the integrity, security, wellbeing, or existence of a person, an entity or an object. A danger may remain without risk, if one knows how to avoid it completely, while a risk always has at its source a danger, which must be identified. For example, in healthcare, a danger might be an infectious pathogen and the risk is the frequency with which an individual develops the corresponding disease.

A **putative risk**, not grounded in scientific or empirical evidence, must also be distinguished from a proven risk. Indeed, a proven risk is never zero. For example, although air travel is the safest form of transportation, the risk of a plane crash is not zero (the probability is estimated to be about 10-7 for current airliners). On the other hand, a putative risk can become zero. For example, in 1836 François Arago, famous scientist and mathematician asked authorities to prohibit people from riding trains because he foresaw a major danger for health beyond the speed of 27 km/h and while traversing tunnels (Arago, 1836). He believed, incorrectly, that the human body would not resist the pressures produced.

The introduction of new technologies in our society can generate not only benefits, but also dangers and risks which must be properly assessed and managed. This is frequently the case, even for very relevant and popular technologies. For instance, automotive technologies are widely diffused and appreciated, but, according to the World Health Organization (WHO, 2020), they are the first cause of death for citizens aged 5-29 years and they have an economic impact, mainly in terms of cost for the healthcare systems, of 3% of the gross domestic product worldwide. Ensuring a safer system approach for all road users is one of the main goals of the UN 2030 agenda for Sustainable Development, which requires important innovation on automotive technologies, including AI-based solutions forgiving human errors. Generally speaking, **governing the introduction of new technologies** in our society clearly requires also to elaborate and promote guidelines, policies and regulatory issues to prevent and mitigate potential risks. These examples call for a precautionary approach to evaluate the reality of dangers, and an evaluation of the resulting risk so that unnecessary measures are not taken for nonexistent or very low risks, and appropriate measures are taken to mitigate proven risks.

When we consider AI-based systems in healthcare, the stakes are high because we are dealing with human life. There are potential dangers, for example, an interpretation mistake in medical imagery could lead to a cancerous tumor going unnoticed; a mishandling or lack of security measures for health data could lead to the disclosure of patient personal data. It is therefore important to correctly qualify the actual dangers of specific technical solutions and to accurately evaluate the related risks so that AI systems are deployed for the benefit of patients and society.

The question is: how to define risk indicators that make it possible to identify if there is a risk at all. For example we could ask the "worst case scenario" questions: *Are there catastrophic consequences of a system failure*? And "*To what extent could the system be considered dependable, i.e. capable of mitigating associated risks for users*"? By performing such analyses, the occurrence probabilities of those events leading to system failure have to be considered, and appropriate measures taken to reduce them to safely manage failure implications and prevent failure repetition. This precautionary process can imply *e.g.*, system redesign, different use protocols, clearer interfaces, user training, and assessment of user capabilities, These measures are classical in critical applications.

The notion of "high" vs. "low" risk underlies a kind of threshold, under which risk could become acceptable. However, using such a binary scale (high/low) might be too limited to express risk impact diversity. A risk scale expressing damage intensity should be *multidimensional*, accounting for different values that could be at risk. For example, data privacy, physical integrity, physical wellbeing, moral impact. In each dimension, risk could be evaluated taking into account several parameters such as

patient context, *e.g.*, age, lifetime expectancy; medical history; healthcare general context, *e.g.*, availability of means or equipment, or of alternative treatments; impact on the healthcare system itself. Risk should also be considered over time, to assess short term and longer term impacts.
Finally, the scale should be a continuum to avoid arbitrary and *a priori* thresholds. This implies a degree of complexity that would be difficult to capture with a binary "high/low" scale.

Dimensions of risk / Influencing factors	Physical integrity	Physical wellbeing	Mental wellbeing	Privacy & intimacy	Agency & autonomy
Personal context					
Personal health condition					
Personal health history					
Age and lifetime expectancy					
Care means availability					

Figure 1. Example of dimensions of risk (or at risk) and factors that influence them. The time dimension is also to be considered due to cumulative effects

The danger/risk analysis assessment and risk mitigation should be performed from the onset of system specification, and continually, after deployment throughout the system's life cycle, taking into account standards and certification processes.

A sound methodology should be developed to correctly make these evaluations and to mitigate risk. In the domain of software engineering, solid concepts and methodologies have been proposed to deal with the dependability or resilience of software systems. Dependability (Avizienis, 2004), defined as "*Delivery of service that can justifiably be trusted*" has several attributes, including system availability (readiness for correct service), reliability (continuity of correct service) and safety (absence of catastrophic consequences on the user(s) and the environment). "Justifiably" means that there is a grounded and proven assessment of these properties. The notion of danger underlies catastrophic failure consequences. Limiting consequences of task failure includes verification and validation techniques, such as error detection and recovery mechanisms, model checking, detection of incorrect or incomplete system knowledge, and resilience to unexpected changes due to environment or system

dynamics. There are means to reach these objectives, such as software system design diversity, redundancy, as well as software architectures enabling system state assessment for decision-making in order to produce error-free results.

This last step may however be performed by a human specialist for example, and requires a specific protocol. For example, the ALTAI (European Commission, 2020b) could be a guide here. This raises issues related to the organization and governance of the healthcare system, and not merely of a piece of software providing a given service. The combination of risk assessment and decision-making is actually a source of complexity, because there is a cost in making the systems fail-safe. Eliminating dangers, i.e., reducing risks while keeping benefits might indeed incur important investments in time and finances as we can learn from the aviation industry for example - hence the 10-7 probability of an airliner crash. But we can also see this approach in healthcare, especially in the pharmaceutical industry and for the design of medical devices. AI systems add a level of difficulty when they are based on learning methods, which are opaque and as such challenge classical verification and validation techniques. A whole field of research currently addresses transparency and explainability issues of AI systems (Barredo Arrieta, 2020). Also, health authorities have already issued guidelines for assessing medical devices which include AI-based systems, *e.g.*, in France (Higher Health Authority, 2020).

4. Case studies

In order to discuss a risk-based approach, we analyze next two case studies that illustrate the use of AI systems in healthcare in order to identify where the implementation of AI systems might bring potential benefits (improvements to treatment outcomes and diagnostic accuracy, healthcare system efficiency, etc.) and to highlight ethical issues, specifically the seven key requirements for trustworthy AI identified by the HLEG-AI (which are the basis for the ALTAI), when implementing AI technologies in the healthcare sector considering a risk evaluation approach as defined in the EU Whitepaper.

1. AI systems for patient triage and prioritization, also dealing with crisis situations such as the COVID pandemic;
2. AI systems for diagnosis.

Case 1. Patient triage and prioritization

When the waiting list of patients is quite long, and the diversity and urgency of healthcare that is sought is multifaceted, an AI system can help to compensate for the lack of adequate personnel to deal with the flow of patients. Recommendations from

an AI healthcare system can help to thoroughly analyze patients' healthcare records in combination with their presenting symptoms.

Based on these factors, the healthcare staff will be able to prioritise and treat those with the most urgent needs. Note also that emotion detection could be used in such systems, which would raise additional ethical issues which are beyond the scope of this paper (see (Grandjean, 2008; Greene, 2019)).

Patient triage and prioritization can be done through an AI system interacting directly with the patients, with the healthcare personnel, or with both. Question and answering systems, or chatbots, which are likely to be the interface to the AI system, will sort the patients to the appropriate level of urgency through a dialogue. The challenge is whether the AI embedded within, or connected to, the chatbot will reliably triage patients to the appropriate level of care. Triage involves a combination of complicating factors--the communication skills of both the AI system and the patient, and the assumption of a factual description of the current symptoms and relevant physical and mental history to mention only two such factors.

A "chatbot" or Artificial Conversational Agent (ACA) is a software system that has natural language processing (NLP) capacities enabling it to enter in a dialogue with a user through a keyboard or a voice recognition and synthesis systems and could also use a visual avatar. One of the first such systems was ELIZA developed by Joseph Weizenbaum at MIT in the mid-1960s, which was based on using keywords and scripts. Interestingly, ELIZA's scripts were based on reformulating user inputs as questions to her/him in a way resembling the communication strategies of Rogerian psychotherapists (Weizenbaum, 1966).

Some of the most known and popular chatbots today are commercial systems such as Amazon's Alexa, Google Home, or Apple's SIRI, connected to the Internet and thus able to access considerable data to answer questions or to conduct e-commerce. Chatbots are also integrated in several specific systems, such as GPS car route planning or queries for travel companies on their websites. Some systems, such as those mentioned above, include a learning capability, see (Kim, 2018), enabling them to improve their response according to new data, previous choices made by the user, or exploiting inputs from other users.

General ethical issues with Chatbots

There are several ethical issues related to developing and using chatbots, and a few of them can be exacerbated when healthcare becomes the application domain. To list but a few:

- *The users might not be aware that they are actually interacting with a computer program and not a human being.*
- *The chatbot's voice and/or appearance might have a specific tone or aspect that might influence user behavior.*
- *The chatbot's behavior will be based on algorithms, which might include AI and learning capabilities, and on a variety of data. Similarly to all such systems, the data may be biased and the chatbot language or behavior as well.*
- *Like all algorithms, including AI systems, a chatbot lacks semantics and does not actually understand what it is doing or what consequences its outputs might have on humans.*

Risk assessment for chatbots in healthcare

Chatbots are also used in healthcare, e.g., in psychiatry (Philip, 2020). In a healthcare context, one must distinguish the "operator" i.e., the medical professional (or organization), which deploys the chatbot, from the "user", the person who is going to actually interact with it through Q&A, from the device manufacturer. The operator is generally not aware about the internal workings of the system, but knows how to use it and what to expect from it. The user is often totally ignorant of the underlying algorithms of the chatbot and its capacities. There are different issues to consider from these two perspectives.

The consequences of chatbot advice or decisions might be severe for the user. One main issue is related to the fact that the chatbot ignores the general context in which it is used, and can only use specific information about patient condition and possibly their medical data. This does not mean the decisions can be wrong. On the contrary, sometimes, and often, the decisions are correct and suggest that the system has been well designed and trained. However, the risk related to wrong decisions remains high because of the consequences for the health of the patients. This has to be acknowledged as a factor for deploying chatbots and evaluating their conclusions by the operator.

Furthermore, correct decisions will tend to increase operators' confidence and trust in the system, perhaps leading them to not question the triage decisions over time.

The chatbot might influence the user through the form and content of its questions and answers, thereby inducing a bias in the user's behavior, that may, in turn, produce a bias in the chatbot decisions. In the instance of an incorrect triage decision, that could prove catastrophic.

The chatbot might not be able to say "I don't know" unless it's explicitly programmed to do so, and might persist in forcing the dialogue to acquire additional data, orienting users' answers. This might produce inappropriate concluding decisions.

Trustworthy AI elements

Almost all of the seven requirements for "Trustworthy AI" need to be considered in evaluating the chatbot AI system, given the ethical questions raised above.

1. Agency: Patients (users) should be informed that they are dealing with a machine (see transparency) and should have the possibility to opt-out and to access a human. Operators should be able to assess and validate the results of the chatbot decisions through metrics (e.g., confidence, performance, explainability).
2. Technical robustness and safety: Chatbots should be verified and validated by certification bodies or trusted third parties.
3. Privacy and data governance: Data collected by the chatbots and underlying platform should respect and comply with general regulations as well as health data sensitivity (anonymity, proportionality, purpose of use, storage and access) as recognized by the GDPR (EU 2016/679).
4. Transparency: The patients should be clearly informed that they are conversing with a chatbot and not with a human through an interface. The purpose of using a chatbot should be stated. The professional operator should be informed about the system's decision process
5. Diversity: The chatbot interface (voice, visual appearance, attitude) should be as neutral as possible and its makers should not try to give the image of a human in its visual appearance to avoid confusion. Issues of diversity are relevant, specifically where facial recognition and emotion detection are utilized for patient sorting. As documented widely, certain ethnic groups can suffer erroneous facial recognition detection, such as populations of color (Grother, 2019). These false-positives might include in addition incorrect emotion detection, which jeopardizes the entire concept of fair patient sorting based on real-time behavioral responses including emotion, etc.
6. Accountability: Accountability and liability must remain with human beings (designers, operators, users, etc.) and not on the machine itself. Indeed, AI systems should not have a legal personality.

It is possible to examine the risk assessment process more thoroughly by applying the ALTAI to one of the trustworthy requirements. The fourth guideline for trustworthy AI addresses transparency, which encompasses three elements: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system. In developing an approach to assessing risk, the HLEG posed some illustrative questions that operators and users of AI in healthcare might employ to identify and mitigate risk (European Commission, 2020b). Their discussion is worth excerpting in the next few paragraphs.

Traceability

This subsection helps to self-assess whether the processes of the development of the AI system, i.e. the data and processes that yield the AI system's decisions, is properly documented to allow for traceability, increase transparency and, ultimately, build trust in AI in society.

- Did you put in place measures that address the traceability of the AI system during its entire lifecycle?
  - ◊ Did you put in place measures to continuously assess the quality of the input data to the AI system?
  - ◊ Can you trace back which data was used by the AI system to make certain decision(s) or recommendation(s)?
  - ◊ Can you trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system?
  - ◊ Did you put in place measures to continuously assess the quality of the output(s) of the AI system?
  - ◊ Did you put adequate logging practices in place to record the decision(s) or recommendation(s) of the AI system?

This could take the form of a standard automated quality assessment of data input: quantifying missing values and gaps in the data; exploring breaks in the data supply; detecting when data is insufficient for a task; detecting when the input data is erroneous, incorrect, inaccurate or mismatched in format – e.g., a sensor is not working properly or health records are not recorded properly. A concrete example is sensor calibration: the process which aims to check and ultimately improve sensor performance by removing missing or otherwise inaccurate values (called structural errors) in sensor outputs. This could take the form of a standard automated quality assessment of AI output: e.g., predicted scores are within expected ranges; anomalies in output are detected and input data leading to the anomaly detected and corrected.

Explainability

Assessing the explainability of the AI system is a second element of trustworthiness. This element refers to the ability to explain both the technical processes of the AI system and the reasoning behind the decisions or predictions that the AI system makes. Explainability is crucial for building and maintaining users' trust in AI systems. AI driven decisions – to the extent possible – must be explained to and understood by those directly and indirectly affected, in order to allow for contesting of such decisions. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contribute that) is not always possible. These cases are referred to as 'black boxes' (Castelvecchi, 2016) and require special attention. In those circumstances, other explainability measures (e.g. traceability, auditability and transparent communication on the AI system's capabilities) may be required, provided

that the AI system as a whole respects fundamental rights. The degree to which explainability is needed - which depends on whom it is intended to (Barredo Arrieta, 2020) - depends on the context and the severity of the consequences of erroneous or otherwise inaccurate output to human life (European Commission, 2020a).

- Did you explain the decision(s) of the AI system to the users?
- Do you continuously survey the users if they understand the decision(s) of the AI system?

Communication

This subsection helps to self-assess whether the AI system's capabilities and limitations have been communicated to the users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy as well as its limitations.

- In cases of interactive AI systems (e.g., chatbots, robo-lawyers), do you communicate to users that they are interacting with an AI system instead of a human?
- Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?
  - ◊ o Did you communicate the benefits of the AI system to users?
  - ◊ o Did you communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/ or error rates?
  - ◊ o Did you provide appropriate training material and disclaimers to users on how to
  - ◊ adequately use the AI system?

Case 2. AI prediction (...and outcome/diagnosis reassessment)

AI has the possibility for greatly changing healthcare – from cost savings through more efficient healthcare, preventing physician burnout by lessening administrative tasks and increasing direct patient care - the use of artificial intelligence in the healthcare environment will be vast and the potential improvements immense. Specifically, the ability for AI to 1) predict onset of health conditions leading to proactive healthcare interventions, or 2) through reaffirming or rejecting physician diagnoses, or finally 3) assisting in patient healthcare record management, can ultimately result in saving lives or increased patient quality of life.

One scenario where AI can improve healthcare outcomes is by predicting onset of health conditions/diagnoses, leading to capabilities to proactively manage healthcare. For example, if a patient is genetically predisposed to cancer, AI can be utilized in personalized medicine to flag the patient's risk of cancer, and recommend healthcare

interventions for the patient. These scenarios are not hypothetical, but could become reality in Denmark and Estonia. In Estonia, for example, there already exists precision prevention for breast cancer or cardiovascular diseases, this is enabled by the nation's biobank (Milani et al, 2015), which currently houses genetic information for 5% of the population. The use of AI could increase the speed of identifying citizens/patients facing these potential diagnoses. Use of AI in routine patient diagnosis could also assist in basic research into disease including classification, prognosis, and treatment.

A second scenario where AI can benefit both healthcare providers and patients is the ability to reconfirm or "correct" healthcare findings (i.e., tests, diagnoses) - similar to the partnership between the UK's NHS and Google's Deepmind AI (King, 2019), healthcare institutions might utilize AI to verify physician findings. In scenarios of telemedicine, such as video meetings, the AI might verify a small sample of digital healthcare meetings to reaffirm that telemedicine is achieving similar, or on-par, results found in face-to-face healthcare.

A third scenario entails AI assisting in patient health record management. For example, after a patient visits their physician, the resulting notes of the visit can be automatically transcribed and added to the patient's health records. The AI can also determine the next appropriate steps for the patient, as well as schedule appropriate follow-up visit communications. The AI system provides advice about the patient condition and further actions to be taken by the physician. In essence, the AI assists the physician by managing the healthcare record.

Given all these benefits from the adoption and implementation of AI in healthcare, there also come corresponding risks and consequences arising from AI in healthcare.

Risks analysis for AI predictions and outcome assessment in healthcare

Various types of risk exist.

Some risks pertain to the quality of the data. AI could potentially both predict and reaffirm health diagnoses. However, some uses of AI to reach medical diagnoses and recommendations may be flawed; there will be a need for ongoing research and transparency (Antun, 2020).

Issues related to poor, erroneous, or incomplete patient data are significant when AI is utilized for maintaining patient records. Training data may be poor, for example, failing to adequately reflect patient groups, may not adequately reflect variations in record keeping (Panch et al., 2019).

Healthcare meetings are complex human interactions and an AI system is likely to miss many elements of human importance. Subtle biases may be incorporated in data recorded in health records which the AI may a) fail to notice and hence reproduce or b) be able to detect and address (Char et al., 2018). Patients' medical data might be located in different electronic health record systems (EHRs enable patient data to be digitally accessed in one central file, allowing stakeholders to seamlessly share, access and exchange a patient's health information (Shortliffe, 1999)), or include data from wearables and other sensors - AI might pose a solution for interpreting data from different sources and different classifications of data, however AI-based solutions are early and mainly used in medical research (Lotman & Viigimaa, 2020).

Some risks may arise from the nature of the data required. Verification of results could be based on a variety of indicators including medical outcomes but also patient and physician satisfaction.

Assessing success of such healthcare meetings thus may necessitate facial recognition and emotion detection AI. This could have risks of discrimination and labelling of certain patients. Situations posing certain risks may arise from the combination of human and AI expertise.

Suppose a physician disagrees with all or part of the AI's recommendations - a system may not allow this; conversely a well-functioning system may be overridden. An institution should develop protocols to deal with such situations.

Mitigation of many risks includes attention to issues beyond AI itself, both within and beyond the medical setting. Within a medical setting, some of these concern pre-conditions for the successful use of AI, some concern possible longer-term impacts of its use. For the AI to function effectively in managing patient health records and advising follow-up communication, prior work is needed integrating computing systems across different sectors. Without this, gains may be fragmentary and illusory (Panch et al., 2019).

There is a possible risk of impact on developing physician's skills and learning from clinical experience, which would need to be monitored and addressed. This is also necessary for good communication from physician to patient regarding their condition and recommendations.

There is a risk of focus on certain technology such as AI at the expense of necessary work on other technologies and the importance of the clinics (symptoms and real-life experiences of the patient).

Faster and easier detection of very early disease stages and focus on risk carries benefits but its routine and long-term use also complex questions pertaining to issues such as risk perception and medicalization which may require relevant expertise to address (Featherstone et al., 2020).

Wider economic and legal issues may arise. Possible flagging/screening by insurance companies of an individual's terminal illness before onset, may lead to exclusion of these patients from the healthcare insurance market. Here, the predictive screening capabilities of AI will not result in saving lives or better healthcare outcomes, but in creating patient discrimination and possible surveillance of those with sensitive or undesirable healthcare conditions/disease/diagnoses (HIV, covid19, etc).

A general unknown risk concerns the future of litigation and case law in medical practice from 'bad cases'. What will happen to the current relationship between a patient and individual physicians with the use of AI systems in recommending treatment? (Char et al., 2018).

Trustworthy AI elements

3. Human agency: such use of AI should not unnecessarily override individual medical judgement and autonomy. Protocols for dealing with mismatch between the judgements of physician and of AI systems raise the risk that this may not always function in the best interests of each individual patient, may for instance be guided by fear of litigation, and/or by focus on certain audited risks rather than on other less tangible risks which are not audited. Will there be room for genuine difference of medical opinion (Char et al., 2018)? Further, patient and physician AI education will be needed for humans to fully comprehend the impact of AI in healthcare – without "AI literacy" humans are not able to fully embrace and protect their own agency.

4. Technical robustness and safety: as outlined in the last section, reliance on AI must not be premature. The AI systems deployed in medical care must be both technically robust, and their technical safety ensured through repeated auditing of such systems (Raji et al., 2020). Further, one of the biggest problems facing citizens is the lack of information about the types of data analyzed in AI systems (Vinuesa et al., 2020). Both issues can be partially rectified by mirroring the use public registers of algorithms used in Helsinki and Amsterdam allowing auditing of such AI systems (Johnson, 2020), allowing citizens to identify the databases that trained the model, how individuals utilize the prediction, description of how each algorithm is used, and how bias or risks were assessed in the algorithms.

5. Privacy and data governance: The protection of sensitive health data utilized in AI-assisted healthcare is not only powerful in its ability to deliver targeted, personalized healthcare, but also has significant issues for potential discrimination or surveillance of patients. Data should remain subjected to GDPR rules which should be strictly applied. Healthcare institutions are based

on trust – trust in the physician, trust in the healthcare institution, and trust in the sanctity of patient data. When trust is broken (not an "if", but a "when"), it is key to identify whom's data was breached, which specific data (i.e. diagnoses or prescriptions), and subsequential potential for fraud or discrimination must be minimized with haste. Critically, individuals should be required to provide clear, meaningful consent for use of their data in healthcare making decisions powered by AI systems, which would enable a better data traceability.

6. Transparency: Transparency and explanation to the physician will be needed at a high level. The importance of checking may be highest for patient groups with reduced capacity to understand the involvement of an AI system, such as those with cognitive impairment.

7. Diversity: imposing uniformity on health care records may be counter to nuances needed to accommodate different groups. Conversely greater ease of personalized medicine and diagnosis may assist in fine-tuning diagnosis and treatment for groups whose disease presentation and treatment may differ from the average of the population. Additionally, it benefits all healthcare institutions to ensure that diverse training sets are utilized in order to make public health decisions. In facial recognition systems, we have seen a lack of diversity in training sets where the algorithms resulted in poor identification of individuals of color (Maurer, 2017, Merler et al., 2017) and therefore databases used for training must hold diversity (in the data) as sacrosanct.

8. Societal and environmental wellbeing: increased diagnosis and medicalization can have downsides as well as benefits, including increasing healthcare costs, weighed against increases in preventative health and personalized medicine which may save costs both monetary and personal costs to the patient of unnecessary or delayed treatment. Unknown risks relate to the possible impact on litigation with complex questions for medical professionals, patients, and society as a whole, including risks of increasing litigiousness in medicine. Societal wellbeing can be jeopardized (including public trust) if debacles such as the NHS' "care.data" scandal are not learned from (Vezyridis & Timmons, 2017). Individuals not able to practice informed consent must be protected and prioritized, as well, as AI brings with it many issues of comprehension and public understanding.

9. Accountability: There will be a certain amount that is unknown about how the law might develop in this area so hospital managers and those in charge will have a responsibility to monitor such situations carefully. Individual medical practitioners and patients also need protection and caution as the full implications become apparent.

5. General discussion and conclusions

The amount of information - from databases of increasing diversity, from the proliferation of sensors, and from smartphone-based apps linked to electronic health records, just to mention a few drivers of change - is beyond the capacity of human intellectual processing. For that reason alone, there will be a steadily increasing use of AI applications by medical and health professionals and the organizations in which they work. These applications will improve healthcare, but they also have the potential to introduce new risks from AI for both patients and professionals. However, trust in purpose and in operation is the foundation for the development and adoption of technologies, and AI is no exception.

Given that healthcare is a high risk sector, these additional sources of risk are beginning to be addressed through the development of requirements for trustworthy AI and tools for assessing risk such as the ALTAI. This report has examined the issue of risk and approaches to assessing and managing risk in healthcare. We have illustrated an approach or use case as to how the seven elements of trustworthy AI might be merged with the ALTAI lists developed by the HLEG to assess some of the risks associated with the use of chatbots in a healthcare setting. The primary argument is that asking focused questions about AI in a specific application or setting from the perspectives of operators and users can help to determine risk and trustworthiness. Our purpose is less to provide specific guidance for assessing trustworthiness of AI applications and more to suggest that developers and users need to take responsibility for developing an appropriate assessment process in their particular setting.

In summary, the following findings have been ascertained:

- Complying with a human-centered AI approach can be assessed through the compliance with the 7 key requirements.
- Risk is not binary. There is a multidimensionality in its nature, a continuum in its intensity as well as a time factor. Assessing risk requires identifying the values that are impacted and the degree to which they are.
- Healthcare is by nature a domain of high stakes. It is also a domain in which several factors are interrelated. A hospital procurement policy may impact its ability to cope with emergency situations. Its management of appointments may impact the availability of beds or operation rooms. It is difficult to assign a priori a risk level to such or such application.

Another important issue is the potential relevance of the correct development of trustworthy AI tools to reduce burn-out of healthcare professionals (correlating with adverse events) and medical malpractice (typically correlating with defensive medicine).

AI could really become an operational tool to manage these critical situations, by optimizing the role of human agents and their liability, thanks to decision-support systems and rigorous, standardised process data tracking. This issue could be very relevant in the short term for the healthcare domain, much more than more radically innovative solutions for diagnoses and therapies.

Additionally, education and training programs for health professionals must, in their various curricula, include a substantive discussion of AI, its promise and potential perils, and management of its risks.

Education for healthcare personnel, whether administration or physician/nurses, is a clear priority, too. Because the public struggles to understand the basics of how AI systems operate (Coeckelbergh, 2019), it should be assumed the same for healthcare personnel. In order for healthcare personnel to understand the risks inherent in use and implementation of AI systems in healthcare, they must be knowledgeable about these systems work - how algorithms arrive at specific decisions, how machine learning and big data can make predictive healthcare diagnoses. Educational literacy about AI for healthcare personnel could follow existing gamification models of digital educational training (Mesko et al., 2015).

Making sure that it is patients who benefit the most from the surge of AI health technology remains a key challenge. This will need new approaches in medical education to improve digital literacy, understanding of mathematical modelling, basis of decision theory, and continuous learning about AI of physicians. This should include awareness of biases in data, and how these undermine any claims about how AI models are able to produce objective, neutral results.

Accountability for AI systems in healthcare is also of great concern. The ability to audit healthcare systems is necessary, and could be built on the aforementioned models utilized in Helsinki and Amsterdam (Johnson, 2020), envisioning physician and patient ability to peer into the "black box" for auditability purposes. Further underscoring the need for education, audibility of AI systems is only possible when stakeholders involved in auditing these very systems comprehend the underlying technologies - where physicians do not fully comprehend all processes involved in AI (Diprose et al., 2020). Ethical-by-design healthcare AI needs to better integrate patients' views and values to understand better different realities and kinds of knowledge, including the subjective aspect of illness. Patients' wishes are a crucial measure for anticipating how AI technologies contribute to their health and wellbeing. Engineers and physicians need to work with patients to establish whether the use of AI is an empowered choice. This will need research programs to understand the patient's own relationship with AI. A first step will be education and allow a better patient's literacy. A second step will be

patient's engagement by feeding the dialogue with AI designers. The final step should be patient's empowerment to gain a better health through self-customized AI use.

Diversity for AI systems in healthcare must be also focused by industry and stakeholders from different perspectives, mainly such as:

1) Diversity of the team of the designers and developers of the AI-based solutions. As a minimum requirement the team should be balanced in gender, so to elucidate the wide spectrum of the needs, behavioural, communication and emotional styles which can be very different for male and female healthcare professionals, patients, relatives and all other human agents involved in the application scenarios

2) AI for diversity, i.e. the implementation of AI-based solutions which should cover the above general specifications by fully exploiting AI also for simulating the wide variety of diversity-open application scenarios, e.g. different facial morphologies and colors, different voice languages, expressiveness and accents, different motor behaviours, different cultural and social contexts. Using AI-based simulations can simplify, accelerate and better calibrate the development process so to deliver highly inclusive solutions.

3) Diversity of the sample population in data. The validation of the proposed AI-based solutions must be carried out by recruiting a diverse set of individuals so as to rigorously assess the actual performance when interacting with different\diverse human agents.

The current pandemic has increased examination of issues related to accessibility of healthcare. The implementation of AI systems in healthcare also brings forth issues of access, where we are now faced with scenarios of affordability. For example, if a private company markets an algorithm for detecting early onset of stroke, how can we ensure all governments have equal access and the technology is not out of reach economically? Rapid developments in AI will indeed increase issues of affordability and access. AI can become a key driver for the development of affordable healthcare solutions, optimizing cost-effectiveness, quality and dependability of novel solutions.

Privacy and security of data are important, as well. Machine learning requires massive amounts of data (Hedlund et al., 2020), and healthcare data possess a higher level of sensitivity and risk versus non-healthcare data. The security of these data and protection of patient privacy is imperative - however, we should not assume that GDPR is flexible enough to keep pace with seemingly quick developments in AI.

Perhaps our most important recommendation is that healthcare organizations need to design an explicit process for assessing AI risk and for mitigating that risk for each application of AI they are considering or using. That process must include the professionals, the organizational leadership, the patients and the public.

References

Anderson T. The theory and practice of online learning. Edmonton, Canada: AU Press, 2008.

Antun V, Renna F, Poon C, et al. On instabilities of deep learning in image reconstruction and the potential costs of AI. Proceedings of the National Academy of Sciences 2020: 201907377. DOI: 10.1073/pnas.1907377117.

Arago, François. Speech before the French Parliament (Chambre des Députés), 14 June 1836, cited in Le Monde, 18 June 1954.

Avizienis Algirdas , Laprie Jean-Claude, Randell Brian , and Carl Landwehr. Basic Concepts and Taxonomy of Dependable and Secure Computing. IEEE Transactions On Dependable And Secure Computing, Vol. 1, No. 1, January-march 2004.

Barredo Arrieta Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Volume 58, 2020, Pages 82-115, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2019.12.012.

Bulla 2020. Bulla, Chetan & Parushetti, Chinmay & Teli, Akshata & Aski, Samiksha & Koppad, Sachin. (2020). A Review of AI Based Medical Assistant Chatbot. 2. 1-14. 10.5281/zenodo.3902215

Char DS, Shah NH and Magnus D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. New England Journal of Medicine 2018; 378: 981-983. DOI: 10.1056/NEJMp1714229.

Castelvecchi, Davide. Can we open the black box of AI? Nature 538, 20–23 (06 October 2016) doi:10.1038/538020a

Coeckelbergh M. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. Science and Engineering Ethics 2019. DOI: 10.1007/s11948-019-00146-8.

Diprose William K. , Nicholas Buist, Ning Hua, Quentin Thurier, George Shand, Reece Robinson, Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator, Journal of the American Medical Informatics Association, Volume 27, Issue 4, April 2020, Pages 592–600, https://doi.org/10.1093/jamia/ocz229

Eaneff, S, Obermeyer, Z and Butte, AJ The case for algorithmic stewardship for artificial intelligence and machine learning technologies. JAMA. Published online September 14, 2020. DOI:10.1001/jama.2020.9371.

European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry (COCIR). Artificial Intelligence in EU Medical Device Legislation. September 2020. https://www.cocir.org/fileadmin/Position_Papers_2020/COCIR_Analysis_on_AI_in_medical_Device_Legislation_-_Sept._2020_-_Final_2.pdf

European Commission 2019a. Independent High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Brussels, 8.04.2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

European Commission 2019b. Independent High-Level Expert Group on Artificial Intelligence. Policy and Investments Recommendations. Brussels. 26.06.2019. https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendationstrustworthy- artificial-intelligence

European Commission 2020a. White Paper On Artificial Intelligence - A European approach to excellence and trust. Brussels, 19.2.2020. COM(2020) 65 final.

https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligencefeb2020_en.pdf

European Commission. 2020b. Independent High-Level Expert Group on Artificial Intelligence. The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. Brussels, 17.9.2020. https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificialintelligence-altai-self-assessment

Featherstone K, Atkinson P, Bharadwaj A, et al. Risky Relations: Family, Kinship and the New Genetics. New York, NY: Taylor & Francis, 2020.

Floridi Luciano, Cowls Josh, Beltrametti Monica, Chatila Raja, Chazerand Patrice, Dignum Virginia, Luetge Christoph, Madelin Robert, Pagallo Ugo, Rossi Francesca, Schafer Burkhard Valcke Peggy and Vayena Effy. AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, December 2018. Available at SSRN: https://ssrn.com/abstract=3284141

Grother Patrick, Mei Ngan, Kayee Hanaoka. Face Recognition Vendor Test (FRVT). Part 3: Demographic Effects. NISTIR 8280. Information Access Division Information Technology Laboratory, National Institute of Standards and Technology (NIST), 2019. Available at https://doi.org/10.6028/NIST.IR.8280

Grandjean Nathalie, Matthieu Cornélis and Claire Lobet-Maris. Sociological and Ethical Issues in Facial Recognition Systems: Exploring the Possibilities for Improved Critical Assessments of Technologies? 10th IEEE International Symposium on Multimedia, 2008.

Greene Gretchen. The Ethics of AI and Emotional Intelligence. Partnership on AI, 2019. https://www.partnershiponai.org/the-ethics-of-ai-and-emotional-intelligence/

Hedlund, J., Eklund, A. & Lundström, C. Key insights in the AIDA community policy on sharing of clinical imaging data for research in Sweden. Sci Data 7, 331 (2020). https://doi.org/10.1038/s41597-020-00674-0

Higher Health Authority, Paris, France, 2020 (in French). https://www.hassante.fr/upload/docs/application/pdf/2016-01/guide_fabricant_2016_01_11_cnedimts_vd.pdf.

Johnson K. Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI. Venture Beat, 2020.

Young-Bum Kim, Dongchan Kim, Anjishnu Kumar, Ruhi Sarikaya. Efficient Large-Scale Domain Classification with Personalized Attention. 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, July 2018.

King, D.. "DeepMind's health team joins Google Health" [Blog]. https://deepmind.com/blog/announcements/deepmind-health-joins-google-health (2019, September).

Lotman EM and Viigimaa M. Digital Health in Cardiology: The Estonian Perspective. Cardiology 2020; 145: 21-26. DOI: 10.1159/000504564.

Maurer D. Face Recognition Technology: DOJ and FBI Need to Take Additional Actions to Ensure Privacy and Accuracy. In: Justice HSa, (ed.). 2017.

McCall B. COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread, The Lancet Digital Health, vol. 2 e166-167, April 2020.

Merler M, Ratha N, Feris RS, et al. Diversity in faces. arXiv preprint arXiv:190110436 2019.

Mesko, B., Győrffy, Z., & Kollár, J. (2015). Digital Literacy in the Medical Curriculum: A Course With Social Media Tools and Gamification. JMIR Medical Education, 1(2), e6. doi:10.2196/mededu.4411

Milani L, Leitsalu L and Metspalu A. An epidemiological perspective of personalized medicine: the Estonian experience. J Intern Med 2015; 277: 188-200. DOI: 10.1111/joim.12320.

Morley, J and Floridi, L. Policymakers must start asking difficult questions on the ethics of AI in healthcare. September 9, 2020. Available at: https://www.publictechnology.net/articles/opinion/policymakers-must-start-asking-difficultquestions-ethics-ai-healthcare. Accessed September 14, 2020.

Morley, J, Machado, CCV, Burr, C, et al. The ethics of AI in Healthcare: A mapping review. Social Science & Medicine 2020; 260: 113172. DOI:10.1016/j.socscimed.2020.113172.

Pagallo Ugo, Aurucci Paola, Casanovas Pompeu, Chatila, Raja, Chazerand Patrice, Dignum Virginia, Luetge Christoph, Madelin Robert, Schafer Burkhard and Valcke, Peggy.. AI4People - On Good AI Governance: 14 Priority Actions, a S.M.A.R.T. Model of Governance, and a Regulatory Toolbox (November 6, 2019). A I 4 P E O P L E, 2019, Available at SSRN: https://ssrn.com/abstract=3486508

Panch T, Mattie H and Celi LA. The "inconvenient truth" about AI in healthcare. npj Digital Medicine 2019; 2: 77. DOI: 10.1038/s41746-019-0155-4.

Philip Pierre, Lucile Dupuy, Marc Auriacombe, Fushia Serre, Etienne de Sevin, Alain Sauteraud, Jean-Arthur Micoulaud-Franchi. Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients.
npj Digit. Med.. 2020-01-07. 3(1)

Raji ID, Smart A, White RN, et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. Fat* '20 2020: 33–44. DOI: 10.1145/3351095.3372873.

Shortliffe, E. H. (1999). The evolution of electronic medical records. Academic Medicine, 74(4), 414-419. Retrieved from https://pdfs.semanticscholar.org/d46d/1c4f5871d3c915d220c7e0350c2c7054583b.pdf [PDF]

Ting Daniel S. W., Yong Liu, Philippe Burlina, Xinxing Xu, Neil M. Bressler and Tien Y. Wong. AI for medical imaging goes deep. Nature Medicine · May 2018 DOI: 10.1038/s41591-018-0029-3

Vezyridis, P., & Timmons, S. (2017). Understanding the care.data conundrum: New information flows for economic growth. Big Data & Society, 4(1), 2053951716688490. doi:10.1177/2053951716688490

Vinuesa R, Azizpour H, Leite I, et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications 2020; 11: 233. DOI: 10.1038/s41467-019-14108-y.

Weizenbaum, J. ELIZA: A Computer Program for the study of Natural Language Communication between Man and Machine, CACM, Vol. 9, Issue 1, January 1966

World Health Organization (WHO), Road Traffic Injuries (February 2020), https://www.who.int/newsroom/fact-sheets/detail/road-trafficinjuries#:~:text=Approximately%201.35%20million%20people%20die,road%20traffic%20crashes%20by%202020

Appendix: 7 Key Requirements for Trustworthy AI (European Commission 2019a).

Human agency and oversight

AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which

can be achieved through human-in-the-loop, human-on-the-loop, and human-incommand approaches

Technical Robustness and safety
AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.

Privacy and data governance

Besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimized access to data.

Transparency

The data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

Diversity, non-discrimination and fairness

Unfair bias must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.

Societal and environmental well-being

AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered.

Accountability

Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate and accessible redress should be ensured.

5
INSURANCE

Authors

Frank McGroarty
Chairman Insurance Committee, AI4People; Professor of Computational Finance and Investment Analytics; Director of Centre for Digital Finance at Southampton Business School, UK

Gianvito Lanzolla2
Professor and Dean at Cass Business School - City, University of London, UK

Nir Vulkan
Chairman Banking & Finance Committee, AI4People; Associate Professor of Business Economics at Saïd Business School, University of Oxford, UK

Paul Jorion
Associate Professor of Ethics, Université Catholique de Lille, France

Patrice Chazerand
Director at DIGITALEUROPE

Rui Manuel Melo Da Silva Ferreira
Chief Data Governance Officer, Zurich Insurance Group (ZIG)

Tilman Hengevoss
Head Public Affairs EMEA Region at Zurich Insurance Group (ZIG)

Xenia Ziouvelou
Innovation Officer and Research Scientist, Institute of Informatics and Telecommunications, National National Centre for Scientific Research Demokritos, & Member of the Scientific Committee on Data Policy and Artificial Intelligence, National Council for Research and Innovation (NCRI), Greece

Executive summary

The impact of AI technology on the Insurance sector seems likely to be in two overarching areas. The first concerns the fair and transparent treatment of consumers. The second relates to risk assessment in underwriting and pricing aspects, in particular when hyper-personalisation of risk assessment is considered.

Algorithms are already being used for every part of the insurance business chain from customer acquisition/retention to risk pricing to claims assessment to investment management. There are obvious customer treatment aspects in the first three areas. Investment management is an important issue for insurers. It is obvious that the capital raised from insurance premia will need to be managed in some way and is usually managed by an investment function within the insurance company who invest in the capital markets. What is perhaps less obvious is that most private sector pensions which are not managed directly by large corporate employers or linked to professional bodies, (i.e. the pension funds of SMEs) are also usually managed by insurance companies. However, investment management overlaps with the focus of the Banking and Finance committee and can be more comprehensively addressed there. That said, given the growing prevalence of AI in investment decisions and in financial markets generally, we feel we should note that financial markets AI poses a systemic sustainability risk to Insurance via asset market crashes and volatility. Insurers can also be entangled with broader financial market risk in another way. In recent decades, some insurers (e.g. AIG) have engaged in reinsurance of structured products and in underwriting financial derivative risk (e.g. credit default swaps). AI algorithms are involved in the risk modelling that determines the pricing of these and in the market behaviour that determines whether their contingent claim is triggered. Again, these issues fit more naturally with the remit of the Banking and Finance committee but we acknowledge the systemic risk to insurers of this AI exposure. "Convergence" triggered by AI adoption across banking and insurance might have a dual effect. On the one hand, it might increase communication and cooperation. On the other hand, it might magnify further the systemic risk described above by enabling "groupthink". In the extreme, this groupthink effect can result in different insurers being overly aligned in the assumptions and models they use to assess and price risk. This could arise through the common use of risk modelling software or data services from single or few providers, or it could arise from insurance professionals coalescing on a single best method or model. Model homogeneity could concentrate and magnify what would otherwise be small risk pricing errors or isolated mistreatments of customers.

"InsurTech" is fast emerging as a distinct area within the wider "FinTech revolution". The latter is causing disruption and a re-think of financial services generally and is strongly assisted by AI technologies. In insurance, AI enables innovations such as micro-

insurance of instantaneous events, bespoke insurance fitting more closely an entity's true risk profile, better bundling and offsetting of an individual's risks, increased operational efficiency leading to cost reductions and greater demographic penetration of insurance services, benefiting the individual and society in general. It also opens the door to "crowd-insurance" which could compete with traditional insurers in a similar way to how crowd-funding offers businesses an alternative way of sourcing capital to banks and capital markets. However, these benefits, which are greatly helped by the use of AI, are also accompanied by the risks introduced in the preceding paragraph.

This committee would also like to highlight the opportunities and risks related to Big-tech – the likes of Google, Apple, Ant Financial - entering into the insurance space. The unparalleled AI and Cloud capabilities of such companies combined with their relatively poorer knowledge of the insurance legacy domains might interact to create insurance policies whose (un-)intended consequences are difficult to predict. Asymmetric access to customer behavioral insights between Big Tech and the insurance industry may also tilt the playing field. Finally, asymmetries in data storage and processing capacity may skew the rules when it comes to enforcing safety and security to a level that matches customers' expectations and trust. As such, this committee suggests that regulators, financial stability authorities and broader stakeholders should systematically assess the actual and potential scenarios triggered by Big-tech's entry in the insurance industry, and act accordingly to maximize societal welfare and ensure fair competition.

An additional consideration for European insurers and policymakers is the prospect of foreign insurers and tech companies arriving in the EU to compete, with ready-to-go InsurTech products that they perfected in their home market, which may not have been subjected to same kinds of consumer protection regulation and legislation that we adhere to in Europe. An example is China's Ping-An Insurance Group, which employs 24,000 software engineers, 800 data scientists, and 180 A.I. specialists, which has invested US$7bn in R&D over the past decade, and which describes itself as a technology-driven finance company[1].

For the purpose of the current report, this committee will restrict itself to an EU, insurance-sector perspective. We believe that the risks of AI to the insurance sector can and should be anticipated, managed and mitigated. If employed carefully and judiciously, AI can avoid perpetuating discrimination and can deliver good, sustainable social and economic outcomes. Broader and better risk management via insurance should aid financial inclusion and enhance financial wellbeing. Conversely, poor implementation of AI could lead individuals to be more cautious in their business and economic activities, to discrimination and ultimately to social exclusion.

1 https://fortune.com/longform/ping-an-big-data/

We highlight the following 5 specific issues:

• ***Technical knowledge gap***:

There is a risk that regulators, senior management and other key decision makers in the Insurance sector, may lack of technical skills to fully understand the effect of technological innovation in their industry. In the past AI used to resemble human thinking and reasoning. But current AI systems such as multi-layered neural networks, are much more opaque. To understand what these systems do and whether they pose a regulatory challenge, for example in terms of discrimination of or deviating from agreed risk levels, regulators need to better understand the inner-workings of the AI system in question. A number of new techniques, which we review in this report, are now available to help them in doing exactly that. These require some quantitative skills and training. It is important that regulators and other key users are trained in these methods and have a more than basic understanding of AI systems and how they work. The same applies to insurance executives who need to understand and be accountable for the workings of their systems. Furthermore, insurers, regulators and policymakers should adjust their recruitment strategies to attract people with these skills.

From a consumer perspective, a basic level of AI literacy will need to be acquired by the general population. Without this there can be no consumer trust in AI tools, which could end up damaging trust in the insurance sector itself. Consumers need sufficient knowledge and confidence to be able to challenge decisions and to understand whether they are fair. And the industry needs to have practices that are sufficiently transparent and explainable to be able to provide consumers with understandable answers about how decisions have been reached.

• ***Social inclusion and fairness***:

Insurance has been framed as a socialization of risk: individual risk is pooled, and insurers compensate their losses with high-risk individuals by profiting from low-risk ones. If we disclose too much of the risk factors, it will be harder and harder to maintain this status quo. As they accumulate more and more data, it will become easier for insurers to quantify individuals' true risk and to identify those customers most likely to make a claim. This poses a potential dilemma for society. If insurers were to solely prioritise profit over all other considerations, they would either avoid those high-risk/high-claim-potential customers, leaving many uninsured, or offer them high prices. A disproportionate number of those likely to find themselves in that situation will be socially disadvantaged and poor, who may not be able to afford the insurance prices offered to them, again resulting in exclusion. The social consequences are obvious but this may also lead to knock-on economic consequences through uninsured people curtailing and suppressing

their own economic activity. On the other hand, conscious behaviour of individuals taking on more risk and assuming that the society will fund the consequences (i.e. free rider problem) goes against the fairness principle. Insurers and regulators need to consider ways of balancing these risks.

• ***Ethical-by-design innovation***:

Engineers building AI systems may lack awareness of the ethical and other wider-implications of the system they are building. They are under increasing pressure to deliver technologies that make faster and more accurate predictions. They may be less aware of the wider impact these systems may have on individual users (e.g. those who are turned down by the algorithm) and society overall. This failure to comprehend the broader societal impact is a trend across the tech industry. It is particularly disturbing in social media where competition for users' attention have created more addictive systems which clearly harm society. Indeed, the AI4People 2020 report on "AI in Media and Technology Sector" found the following: "The sector is more directly user-facing compared to other sectors such as the energy or automotive sector, with, for example, social media platforms being essential for social interaction and information sharing. This means that people might peg the confidence they should have in digital technology to how much they can trust social media platforms. However, at the same time, the MTS offers the opportunity to provide AI with a promising front office, by realistically framing doom stories and possibly showcasing the advantages of cutting-edge technology. In that way, people can learn the ropes of empowerment in an environment which is more familiar, or less forbidding than anything related to health or mobility."

We propose regular training for engineering and other technical staff working in insurance to highlight, among other things, the EU seven requirements for Ethical AI. Also, while we note the insurance industry's clear awareness of the potential for emerging technology (personalized apps and IoT devices) to 'nudge' consumers towards better/healthier behaviours, we advocate the use of sandbox frameworks to explore unintended behaviours of these AI systems prior to deployment. We advise promoting the principal of ethical-by-design. However, it may be more pragmatic to emphasise the more measurable concepts of environmentally-friendly-by-design and socially-beneficial-by-design. This will require an intelligent approach to incentivisation and to regulation, including self-regulation and co-regulation, in order to achieve trustworthy AI in insurance, with the desired social and green outcomes, without stifling innovation. In addition, we advocate that the industry adopts common information standards, to make insurers' systems and platforms interoperable, and data easily transferable[2].

2 EIRA (see TOGAF (2017)) could be a good starting point for developing insurance industry system interoperability.

This is in the interest of customers, including business and employer customers, who are best served by being able to compare competitor insurance products with existing and eases transfers of customer data between old and new insurance providers. It also facilitates data aggregation for regulators to assess systemic risks.

• ***Humans always in charge***:

As AI systems get better, the risk increases of insurers becoming more reliant and less likely to challenge the decision made by these systems. Where possible, AI systems should not replace human decision making but rather should be viewed as an additional tool in the decision-making process. Radiology provides a very good example: Here AI systems are significantly better than humans in identifying abnormal patterns in images. There were some about 6 or 7 years ago who predicted that these AI systems would replace radiologists all together. That clearly did not happen, and these systems are now being used widely by the professionals to make better decisions. We believe this is the model insurers should follow, rather than replacing humans with machines. There is a precedent for this in finance, in algorithmic trading which is a form of AI that has been in existence since the late 1980s. Here algorithmic models which take advantage of momentum (or trends) in financial markets have been widely used for decades. We know that these strategies tend to do well in normal market conditions. However in some extreme market conditions (but not all) they can crash badly often losing in matter of days much of the gains made in the preceding months[3] and often interact with other autonomous algorithms in unanticipated ways. This phenomenon is known as "momentum crashes". Hedge funds which follow these models blindly suffer. However, funds which use oversight risk management strategies where final decisions on leverage levels are made by individuals, have done better and suffered less. AI systems used by insurers are more likely to be customer-facing than market facing. However, we think the principle of human oversight of AI decision-making systems remains the same.

To keep complacency in check, it may be worth noting here some of the proposals of the "Governance Innovation" report released by Japan METI[4] at the OECD (in January 2019). For instance, this report proposes goal-based regulation "[which] will enable more flexible policymaking based on the [nature of the] risk. In other words, by requiring business to implement reasonable measures based on the nature/size of the business and the impact on consumers, policies can be implemented flexibly, i.e. businesses with a strong influence on society are required to protect the interests of law more prudently, while businesses with smaller risks

3 See https://www.aqr.com/Insights/Research/Journal-Article/Momentum-Crashes for more details
4 https://www.meti.go.jp/press/2020/07/20200713001/20200713001-2.pdf p. 42

are allowed to take a less onerous compliance approach. In this age where changes are rapidly taking place and the need for revolutionary innovation is increasing, setting the rules in advance could frequently create adverse effects, therefore it would be better to have goal-based regulations as a basic policy."

• ***Data risks***:

Sophisticated AI systems are "data hungry" in that performance and accuracy depend on being able to run on very large data sets. In fact, the amount of data the average AI system requires seems to increase exponentially. As more interactions between individuals and insurers move online, new kinds of data is being collected on everything from mouse moves, to how long user stayed at what page, to how many clicks she made, to ambient conversations, to location information, etc.. This opens up issues of privacy and data ownership, and the laws and regulations in place to safeguard these. A further complication is posed by the fact that these data sets are increasingly being stored on cloud services. Recent incidents suggest this is a real and present danger which all varieties of financial institution must act to guard against.

Having outlined these challenges, we note that insurance is already a highly regulated sector. Many of issues covered in the media such on so called "AI bias," where algorithms learn to discriminate against certain groups (e.g. gender, race), are already addressed by existing regulation. Furthermore, additional layers of regulation will increase the barriers to entry for start-ups and smaller firms. This can result in unfair advantages for deep-pocketed, transnational insurance companies and tech giants who are now entering this space, which could be detrimental to competition and to consumer choice.

We hope that the steps highlighted in this document can address and mitigate the risks posed by AI in Insurance, avoiding too much additional regulation and the costs and inefficiencies that are associated with it. Complex and convoluted regulation can result in the opposite outcome to the one intended, especially in relation to consumer protection. This is because complex regulation can only be navigated by those who can afford the lawyers to navigate it. This can cause concentration of market share in a small number of providers, reducing consumer choice and increasing systemic risk.

We note that the financial crisis of 2007-08 began with mortgages, arguably the most regulated part of finance at the time. At that time, the creation of sophisticated products, involving banks and insurers, on the back of mortgages resulted in the excess risk which culminated in the crisis. We hope to avoid a repeat of this in coming years where AI technologies will undoubtedly enable the creation of additional, unnecessary risks in the financial system. The steps we highlighted, especially around the technical training for regulators, we hope can serve as a big step in the right direction.

In the rest of this document we consider the impact of AI technologies on the Insurance sector in light of the EU 7 key requirements for trustworthy AI. For the purpose of the Insurance sector, we have merged some of them together as they are addressed by the same recommendations made in the report. These are:

Requirements 1, 4 & 7: Human agency, oversight, transparency and accountability
Requirement 2: Technical robustness & safety
Requirement 3: Privacy & data governance
Requirement 5: Diversity, non-discrimination and fairness
Requirement 6: Societal-environmental well-being

For each we provide use cases, a review of the research and literature in both academia and by insurers and regulators before making our recommendations. We suggest a novel system of labelling to indicate trustworthiness of AI applications. We evaluate the various regulatory framework options and we introduce a classification framework for evaluating the necessity for regulation based on both the risk and the consequential dependence of AI decision outputs.

1. Introduction

The aim of this paper is to create a Good AI Global Framework for the Insurance sector with the concrete objectives of: (a) considering the impact that the 7 Key Requirements will have within the sector, and (b) establishing concrete and practical steps (concrete actionable recommendations and high-level guidelines for distinct stakeholder segments) that the insurance sector must take to be compliant with respect to the 7 Key Requirements for a Trustworthy AI.

The insurance sector is defined here in the broadest way possible, meaning a wide range of activities covering pensions and asset management as substantial key activities that insurance companies do, alongside more obvious insurance functions.

Furthermore, this paper will not discuss insurance ethics in general, but rather the ethics of the AI-driven insurance. It aims to distinguish between high-level guidelines and concrete actionable recommendations for firms.

2. The impact of ai for the insurance sector

Insurance sector overview and key stakeholder segments (value-chain)

We can think of the insurance stakeholder model as having the insurer at the centre. Around the centre sit a number of other entities: premium payers, insurance beneficiaries, capital markets, insurance company staff, insurance company shareholders, regulators and government. Premium payers pay into the insurer in order to get insurance cover. Beneficiaries receive a payout in the event of a claim. The premium payer may also be the beneficiary, but this is not necessarily the case. The capital markets are another stakeholder because this is where the insurers invest the premiums in order to maximise their value so as to be able to pay future claims and also to sustain the insurance companies and reward their own shareholders. Insurance regulators are responsible for the robust and fair operation of the insurance industry. Government has an interest in the insurance industry that goes beyond policymaking. In many instances, government may end up having to pay if an insurance industry solution fails to deliver.

Data processing and data-led statistical analysis has always been the core of the business of insurance undertakings. Digitisation enables the emergence of new types of data, which combined with increasingly powerful IT tools, algorithms and information systems, facilitate more predictive, descriptive and prescriptive analytical processes. AI tools and algorithms can discover and test hypotheses, make decisions automatically and access previously inaccessible datasets, such as unstructured data from pictures, videos and audios.

Table 1: The Insurance Value Chain: Past and Future perspectives

Insurance Value Chain

	Product development	Pricing & Underwriting	Sales & Distribution	Operations	Claims
	Product specifications, rate calculation, market launch	*Actuarial analyses, risk selection, reinsurance*	*Marketing, sales, distribution, channel management*	*Customer management/ engagement*	*Claims/Claims management, billing/collection*
Past	One size fits all policies, high deductibles	Limited data & historical regression analysis	In-person agents, call centers, basic online functionality	Manual, on-premise, administration processing	Manual, on-premise, claims processing
Future	Data & AI inform future product needs	Evaluate risk & price more granularly, high degree of service bunding	Enable personalised Offerings & strengthen value proposition	Increase customer engagement & loyalty & realise churn potential	Improve fraud prevention & detection
Future	Customized, low deductible, per-use, new products made viable	Big Data & Machine-learning predictive analysis	Online, mobile, social, comparison shops	Cloud-based, online, mobile, higher levels of satisfaction	Cloud-based automatic, instantaneous billing

(Extending McKinsey 2020)

Big Data and Artificial Intelligence has significantly impacted the insurance value-chain (Table 1). All value-segments have been transformed by Big Data, AI, and IoT (Internet of Things) among others. According to a recent study by EIOPA (2019), Big Data Analytics (BDA) tools are commonly used by insurance firms in the motor and health insurance segments. Such tools are generally focused on a specific part of the insurance value chain and very few firms make use of them across all their processes. BDA tools are mainly used for pricing and underwriting (35%), claims handling including fraud prevention (30%) and sales and distribution (24%). Furthermore, although the vast majority of companies use in-house developed solutions, there are many others that use "off-the-shelf" solutions from third party service providers and open source (i.e. freely available) tools.

Unique aspects of the insurance sector (social value of insurance, cultural sensitivity of insurance, etc.)

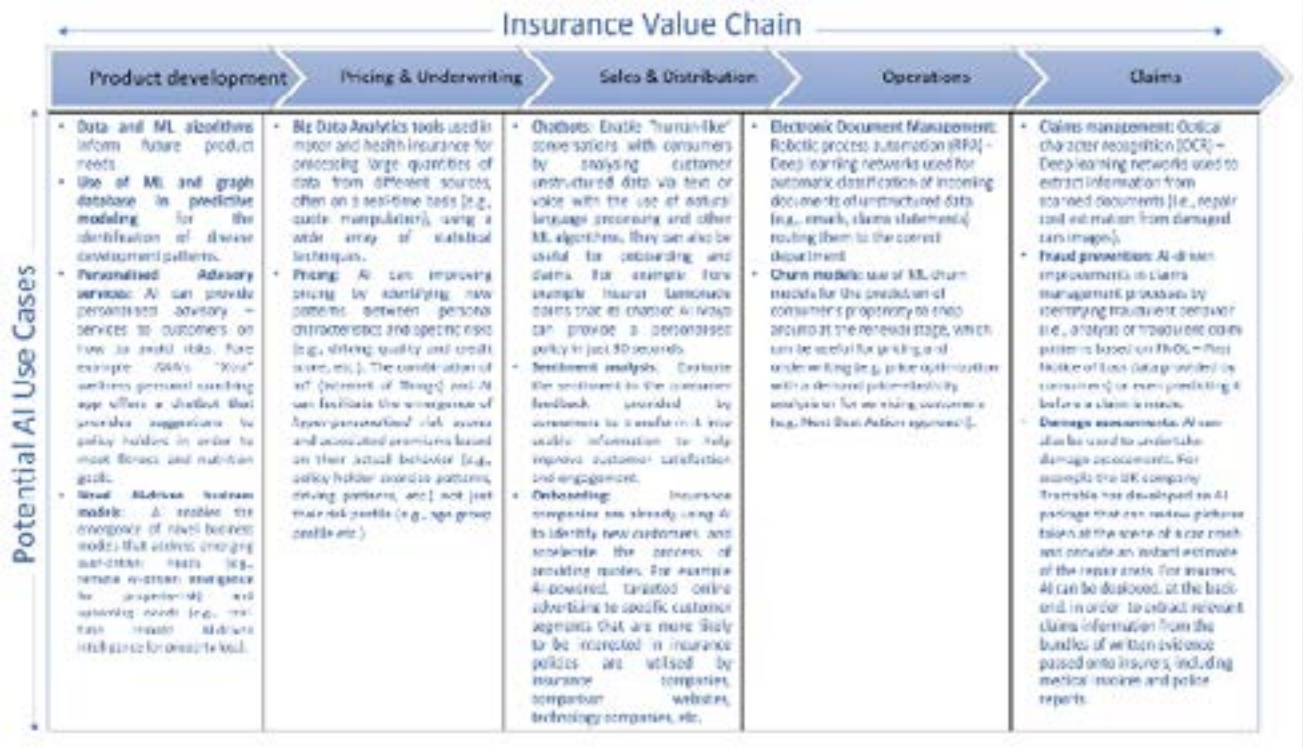
There are social values of insurance that go beyond the immediate mutualisation of individuals' risks. Firstly, more economic activity will take place because insurance is available than when it is not. This is because at least some individuals will moderate their activity in the face of uninsured risk. Secondly, insurance enables some groups to subsidise others in socially beneficial ways - e.g. less vulnerable to more vulnerable, richer to poorer, younger to older, healthy to ill, and among ethnic groups, different

genders, etc. There are economic redistribution aspects to this and there is also the effect of a greater proportion of society being insured which amounts to a positive externality.

Opportunities and Risks of AI for the Insurance Sector

Deployed responsibly and in competitive markets, AI could provide numerous benefits across all areas of the Insurance value chain (see Table 2).

Table 2: AI in Insurance: Potential use cases across the Insurance value chain



(Extending EIOPA BDA Thematic Review)

What are the ethical implications of insurers using AI?

The issue of AI in insurance is the issue of processing big data. "Large societal benefits arise with the potential to reduce risks and increase insurability through the use of vast quantities of data. New approaches to encourage prudent behaviour can be envisaged through big data, thus new technologies allow the role of insurance to evolve from pure risk protection towards risk prediction and prevention. However, the use of big data in insurance raises complex issues and trade-offs with respect to customer privacy, individualisation of products and competition. Assessing these trade-offs

requires complex value judgements, and the way they are addressed leads to different scenarios for the future development of the sector.", Anna Maria D'Hulster, Secretary General, The Geneva Association (2020).

Ethical AI Dilemmas in the Insurance Sector

The insurance sector faces a number of dilemmas. A central challenge is that business optimizing actions do not necessarily align with what is deemed ethically permissible (the trade-off between business optimizing and legitimacy is one of the key trade-offs that AI poses to organizations). For example, although insurers have legitimate reasons to use AI in the way they do, many of these behaviours and/or actions are misaligned with what society in general or some in society find acceptable. In many cases, customers may have diverse views as to what they see as a valuable and ethical use of AI and data processing (CDEI, 2019). According to an online survey that was conducted by Deloitte in 2015 (2,955 insurance customers with motor, home and health policies), 40% of policyholders would allow insurers to track their behaviour and associated data (telematics-based insurance), for a more accurate healthcare insurance premium, as opposed to 49% who disagreed. The percentages for home insurance were 38% and 45% respectively whereas for motor insurance 48% were willing to share their data versus 38% that disagreed. The adoption of telematics-based insurance and the willingness to share data with insurers differs between young and older customers as well. Digitally-savvy young customers (63% of customers between 25-34) were found to be more willing to share their data with home insurers in return for a more accurate premium, than older customers (38%).

Another challenge is that many are convinced that AI will raise concerns for policyholders. This is reinforced by cases where insurers appeared to be creating innovative insurance services without fully considering the ethical consequences. For example, in 2016 Admiral was criticised for attempting to use Facebook posts to analyse the personalities of car owners in order to set the price of their car insurance (Guardian, 2016). Finally, the insurance firm was forced to abandon its plans, within a few hours, as the scheme breached Facebook's privacy rules regarding its users. The scheme was launched later with "reduced functionality" and users could log in to the service with Facebook, without any intrusive attempt to analyse their posts. Although this reversal was welcomed, it raises concerns regarding potential future attempts of other companies trying to use personal data in a similar way. Such encounters create challenges for users that may find it difficult to avoid opting in, as the financial disadvantage of such an action may be quite significant that users are left without any other option than allowing companies to access their data.

3. Use-case analysis regarding the 7 key requirements for trustworthy AI

In considering use-cases for the 7 Key Requirements for the context of the Insurance industry, we found considerable repetition in issues we explore for the various requirements. This was particularly the case for requirements 1, 4 and 7, which, for brevity, we address collectively.

3.1 Req. 1, 4 & 7: Human Agency & Oversight; Transparency; Accountability

REQUIREMENT 1: Human agency and oversight

Including fundamental rights, human agency and human oversight

REQUIREMENT 4: Transparency

Including traceability, explainability and communication

REQUIREMENT 7: Accountability

Including auditability, minimisation and reporting of negative impact, trade-offs and redress

The **relevant fundamental rights that companies within the insurance sector** should consider are:

- **Dignity and non-discrimination**
  - o *"All human beings are born free and equal in dignity and rights."*; *Article 2: "Everyone is entitled to all the rights (...) without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status."* (Universal Declaration of Human Rights – Article 1)
  - o Centre for Data Ethics and Innovation on AI and Personnel Insurance – *"Insurers are prohibited by law from discriminating against customers on the basis of their sex, ethnicity and several other characteristics."*
  - o European Commission – *"The use of AI can affect the values on which the EU is founded and lead to breaches of fundamental rights, including the rights to (...) human dignity, nondiscrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation (...)."*
  - o *EU Gender Directive: The EU Gender Directive prohibits gender-based differentiation of insurance premiums*[5].

5 Council Directive implementing the principle of equal treatment between men and women in the access to and supply of goods and services 2004/113/EC, 13/12/2004.

- Related to dignity, we believe it is the right to an explanation and social accountability: human beings should be entitled to an explanation when an AI system makes a decision about whether they can get insurance and at what price. Ethical Guidelines for Trustworthy AI: "Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process." (page 18). Transparency by design is a frequent feature in AI systems developed by startups that operate in Europe - e.g., Lemonade and Akur8.

  Furthermore, in order to ensure that consumers are not unfairly discriminated, public authorities should independently monitor the data used by insurers and the output of AI systems.

- **Personal autonomy**
  - *"AI could one day be used by insurers to advise customers on how to avoid risks, for example with chatbots suggesting healthy eating and exercise regimes. Some believe behaviour change initiatives like these would impinge on the autonomy of policyholders, while others say they could result in meaningful improvements in people's living standards." (Centre for Data Ethics and Innovation on AI and Personnel Insurance). There are several examples of European start-ups which are using gamification as a proven way to nudge people into behaving. Nevertheless, the jury is still out there on assessing the real net effect of gamification on behaviors.*

- **Personal and societal implications**
  - *"AI is set to make risk assessments more accurate by revealing new predictors of risk. This could result in some groups paying more for their insurance premiums, possibly to the point where products become unaffordable. Yet the opposite may also be true, with AI-powered risk assessments showing individuals to be less risky than they first appear (e.g. some young drivers). If large parts of society become uneconomical to insure, a wider debate will be needed on whether the state should intervene, and if so, on what terms." (Centre for Data Ethics and Innovation on AI and Personnel Insurance)*

- **Personal data and privacy protection**
  - *"While insurers may be tempted to store this data, perhaps in the expectation they will be able to put it to use in future, doing so raises several ethical concerns. One is the threat to people's privacy[6], especially where datasets are at risk of a cyber breach. Another relates to fair compensation [for selling this data]." (Centre for Data Ethics and Innovation on AI and Personnel Insurance)*

6 In the pharma industry, some companies are experimenting with a new "flipped" approach: patients own their data and give permission to companies to use them, on a case-by-case basis. Furthermore, patients can withdraw permission anytime. Permission is often also linked to ad-hoc financial compensation.

- "*Requirements aimed at 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 General Data Protection Regulation and the Law Enforcement Directive regulate these matters." (European Commission)

Insurers have the right to privacy and data protection. Insurance AI tools and systems must respect EU laws on privacy and data protection (The Charter of Fundamental Rights of the European Union, articles 7 (private and family life), 8 (personal data) and 52, (2012/C 326/02), General Data Protection Regulation and the Law Enforcement Directive). Insurers private life should be protected in order to ensure citizen freedom, while at the same time consumers should be able to to access insurance policies that rely on non-intrusive data processing practices.

- **Oversight and accountability**

In order to ensure legal compliance, algorithmic tools in the insurance context should go through, systematic assessment(s). These assessments, including impact and risk assessment, should be conducted before the launch, but also during the whole product's lifecycle. This monitoring and assessment process should be conducted not only by the insurance firms but also by dedicated public authorities.

Accountability

In 2018, a study by the Financial Conduct Authority (FCA)[7] was conducted examining the complexity of General Insurance (RI) pricing models and practices in the UK retail home insurance market.

The study identified a number of issues relating to firms' pricing practices that present the most potential for significant harm and poor outcomes for consumers. Such as that firms did not have appropriate and effective strategies, governance, control and oversight of their pricing practices and activities, and as a consequence they were unable to reliably assess and evidence whether they are treating their customers fairly.

They found that several insurance companies were unable to name a dedicated member of staff who had ownership over their pricing strategy, which could include how AI-led risk assessments influence premiums. As the study notes: "appropriate governance and controls mechanisms are need to be underpinned by clear lines of accountability and responsibility so that firms appropriately consider and evaluate how pricing decisions impact consumer outcomes. This includes considering whether the pricing structure and approach meets our requirements on firms to have due regard to the interests of its customers and to treat them fairly." (FCA, 2018).

7 FCA (2018), "Pricing practices in the retail general insurance sector: Household Insurance", Thematic Review TR18/4, October 2018.

- **Human agency:**

Relevant governance mechanisms for achieving human oversight are: human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. In the insurance sector this implies training staff in the appropriate technical skills and hiring technically competent into areas of the business that were not previously considered as technical support. It also means training people in business ethics and having audit procedures that check periodically whether measures are actually effective. The healthcare industry is experimenting with a new paradigm: "human-in-command", which is considered to be a must whenever sensitive data is at stake. Ultimately, in these models a human gatekeeper validates, or rejects, the AI-based recommendations.

Arguably, human agency in the insurance sector is informed by human agency in sectors that cannot operate without being insured. In transportation, for instance, insurance premiums have long been shaped by stable conditions whereby a pilot would be the ultimate decision-maker in flying a plane, or a driver should be in control of his vehicle at all times lest he might qualify for fines. Nowadays, software may overrule the captain, as illustrated in B737MAX crashes, or Autopilot is behind Tesla's wheel. As a result, human agency in the insurance sector will of necessity reflect radically lower human agency in industries they work with, transportation, health, energy, etc. A possible consequence is for insurance companies to adjust to this trend passively. Another option might have them contribute to the works of the relevant sectors by evaluating early-on the insurability of particular paths considered: insurable scenarios would be deemed fine, those ineligible to insurance would carry a red flag or be abandoned. For instance, will companies which are unable to get a quote by a legacy insurance company decide to design their own insurance– e.g., Tesla for its autonomous car business?

This may be the idea behind the co-creation model presented in METI's "Governance Innovation" report (cited earlier in the report[8]): "Mechanisms and systems for achieving the objective of law differ by company. Further, the shape of cyberspace is invisible, and information needed for evaluation of cyberspace is asymmetrically accumulated in businesses, therefore it is difficult for a third party to determine from the outside how businesses are actually ensuring compliance. Consequently, in order to encourage businesses to achieve sustainable innovation and protect social values set by law at the same time, a co-creation model would be appropriate in which businesses that design/manage cyber-physical architectures to design, disclose and explain their approaches and concepts of compliance to the regulatory authorities or markets, timely receive feedback from them, and continue to evolve."

8 Japan METI at the OECD (2019).

Furthermore, AI adoption across sectors has other systemic consequences on human agency. AI adoption is positively associated to convergence in management attention in firms across industries – including industries which are usually considered as unrelated such as insurance and pharmaceutical (Lei, Lanzolla and Tsanakas, 2020). While convergence in management attention might make communication and cooperation easier across firms and industries, here we want to note some potential pitfalls. First, it might increase "groupthink" and by implication, systemic risk. Second, we note here that convergence might reduce the space for strategic differentiation for insurers thus potentially opening market opportunities from other market spaces, including big tech.

Multi-stakeholder consensus on what constitutes a responsible use of AI and data

The industry should engage with the public in order to reach a consensus on what constitutes a responsible use of AI and data. For example, in relation to personal data processing from social media; a joint decision should be made regarding the conditions under which it is acceptable to process data from social media platforms or to use algorithms to predict people's willingness to pay higher premiums. In UK the Financial Conduct Authority (FCA) has developed a Framework for assessing when price discrimination may be a cause for concern, that is when it can potentially disadvantage some consumers significantly, in particular the most vulnerable and least resilient consumers. The proposed 6 question Framework aims at addressing the issue of unfair pricing practices in retail markets.

In addition, it should also be examined whether tighter controls need to be in place on the use of personal characteristics in pricing. If for example, AI enables insurers to identify high risk characteristics that were not possible before (e.g., chronic health conditions, etc.), then this could in turn result to more people facing unaffordable premiums. As such, "society should have a say in any decision on where to redraw the boundary between acceptable and unacceptable forms of discrimination" (CDEI, 2019). However, in the course of this search for consensus, the industry should proceed with pro-active interventions (rather than re-active) to address obvious harms. Industry-driven measures should be set in order to ensure that AI and data are used for the public and common good. Such measures may range from more accessible privacy notices and data discrimination audits, to industry-wide registers for third party suppliers of data (CDEI, 2019). Furthermore, all these measures should be supported by a sector-wide commitment to transparency.

There needs to be "…. a sector-wide commitment to transparency. Without greater disclosure, insurers will struggle to build trust with customers and regulators will lack the information to design proportionate regulatory responses. ……. [In addition], greater transparency would help to distinguish genuine threats from those that are overstated and would support the development of interventions that are proportionate to the risk in question, thereby allowing responsible innovation to flourish." (Centre for Data Ethics and Innovation on AI and Personnel Insurance)

Explainability / transparency for insurance:

A central issue is being able to explain to consumers how decisions about their individual premiums or claims have been reached. In the UK, the Information Commissioner's Office (ICO) is the regulator with responsibility for upholding information rights in the public interest. They advise on how firms can comply with information and privacy legislation such as GDPR and the Data Protection Act. The ICO recently had an initiative called Project ExplAIn, which aimed to help organisations explain decision outcomes of AI systems[9]. They also developed an Auditing Framework for AI, which helps regulators to evaluate algorithmic fairness.

Other stakeholders also have an interest in transparency. These tend to be more macro-level and systemic in nature. These can include the entire industry being dependent on the same set of assets performing adequately in order to be able to pay future liabilities. Also, if all insurance companies use the same modelling approach, or even buy in the same third-party models, there is a risk that mistakes in these models could be compounded and magnified.

Need for Transparency in Insurance

The need for transparency exhibits variations depending on the type and significance of the AI insurance applications, and the extent to which there has been a change in the decision-making logic and/or the data sources utilised (Figure 1).

9 It is worth noting that we have seen significant progress in recent years in explaining complex AI models, such as SHAP (SHapley Additive exPlanations) values and Local Interpretable Model-Agnostic Explanations (LIME), as well as Optimal Classification Trees to improve accuracy and while addressing interpretability issues.
9 See https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016807c65bf

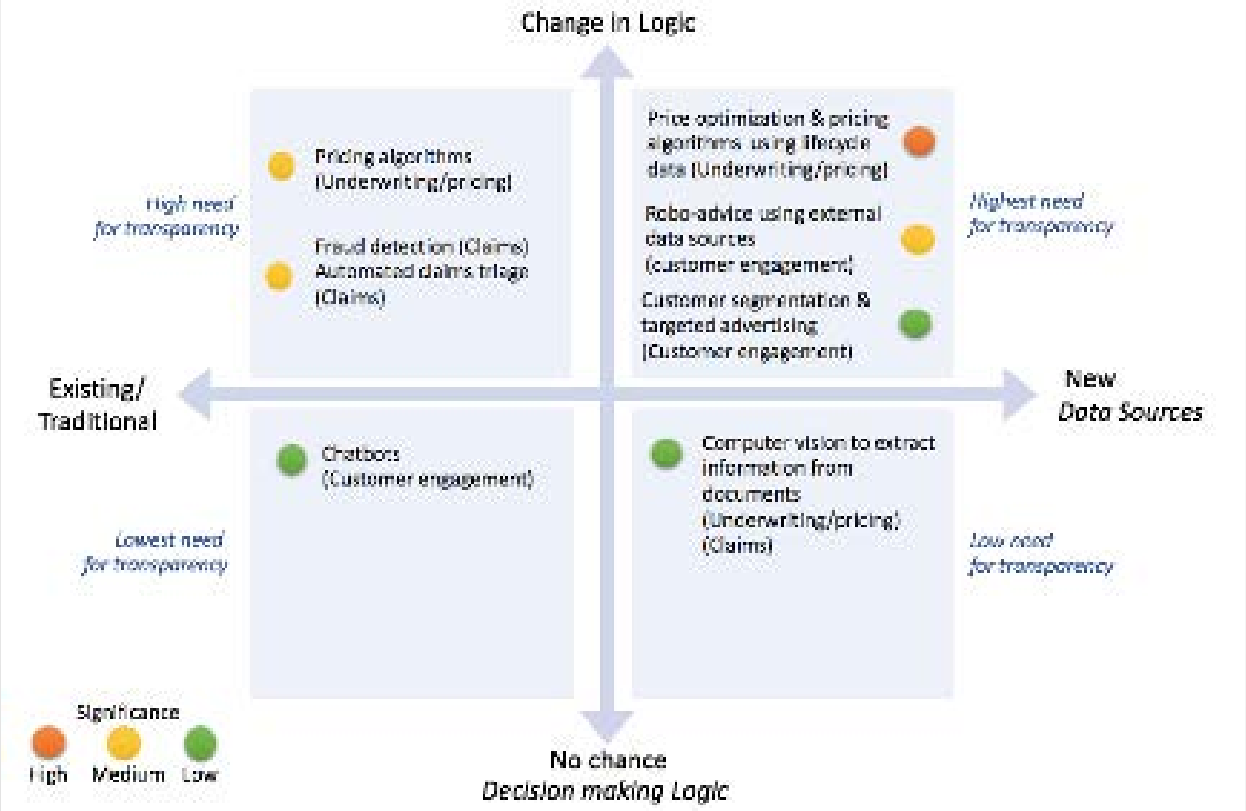


Figure 1: Classification of AI Insurance applications, their significance and the associated need for transparency (extending GA, 2020)

For some use cases relying on complex non-linear models (e.g. deep learning) or on the use of new external data sources explainability might be more difficult and affect customer acceptance. To offset these challenges, some companies are experimenting with algorithms which reverse engineer the decisions from the original algorithms, without limiting the complexity of the original algorithms themselves. These explanatory algorithms may not reveal all aspects of the decision process, yet they might shed light on some "patterns" underpinning the decision. Nevertheless, although this is possible, it is still unclear whether these "reverse engineering algorithms will be sufficient for mitigating the explainability challenges.

Existing legal requirements for transparency to (retail) customers should be interpreted for the use of AI in a balanced approach. The customer's right to ask for human (re-)validation of the correctness of the algorithm-generated result and respective reasoning for the correctness of the response should address customer acceptance issues.

Providing an explanation is particularly important when a decision has a significant impact on the affected individual. Therefore, the degree to which explainability is needed is highly dependent on the context and the severity of the consequences when an output is erroneous or otherwise inaccurate (European Commission 2019).

Interpretability of algorithmic outcomes is important to provide meaningful explanation to affected individuals; and is also indispensable for assessing the performance of AI systems and for their continuous improvement, and thus for sound data science.

The implementation of interpretable models is encouraged, in particular if their outcomes have a significant impact on customers. When used for risk selection and pricing, trust in AI systems can be fostered by using data sources that are related to the insured risk in a way which is intuitively understandable to customers. Where it is difficult to explain algorithmic outcomes in an understandable way to consumers, there are other measures that can foster customer trust, such as *traceability*, *auditability* and *transparent* communication about a system's capabilities (European Commission, 2019).

Measures of transparency:

- Keeping records and data: records of training data sets (selection process, characteristics of date); documentation on the programming and training methodologies; disclosure upon request in particular for inspection by competent authorities (European Commission)
- Internal guidelines and policies: Insurers can mitigate the risk by developing and implementing internal guidelines and policies to ensure a consistent approach to the transparency and explainability of algorithmic outcomes. Guidelines should help to clarify how the benefits and risks of using AI should be assessed on a case-by-case basis. Actuaries, risk managers, data scientists and data protection officers, product development and digitization managers, innovation executives (just to mention a few key stakeholders) should closely cooperate in the development and implementation of such guidelines and policies.
- Communications, Traceability and Explainability: Humans should be informed when they are interacting with AI systems. If someone wants to challenge a decision or a price, the factors and sequence leading to that decision should be fully traceable and explainable in a way that a lay person could understand.

Un-intended consequences at an implemention level

Transparency has also several potentially negative implications such as "sucking up the system" and transactional behaviours – there is significant evidence of this in the management literature. Or, a blind faith in data governance might hinder the ability of an insurers of modelling what the industry call "emerging risks".

Ethical Algorithm Audits

The need for an ***ethical audit framework for algorithmic development*** and deployment has been proposed by a number of stakeholders from the research and the industrial communities. Such ethical "algorithm audits" (O'Neil, 2016; Larsson et al. 2019) could function as AI auditing mechanisms, ensuring that the moral and ethical issues surrounding the use of AI are being addressed, while identifying potential biases or flaws, depending on the type of industry and globally accepted auditing procedures and standards.

Larsson et al. (2019) suggest a role for ***professional algorithm auditors***, whose job would be to interrogate algorithms in order to ensure they comply with pre-set standards. One example would be an autonomous vehicle algorithm auditor, who could provide simulated traffic scenarios to ensure that the vehicle did not disproportionately increase the risk to pedestrians or cyclists relative to passengers.

The accountability and responsibility of AI systems and their outcomes is a key aspect of trustworthy AI. Hence, necessary mechanisms should be introduced in order to ensure them in the insurance sector among others. As the backbone of privacy and data protection regulation at a global level, the principle of accountability is reflected in Europe's General Data Protection Regulation (EU), 2016/679 (GDPR), requiring data controllers to implement appropriate technical and organisational measures. Core elements of *organisational accountability* include proportionate procedures, top-level commitment, risk assessment, due diligence, communication and training, monitoring and review (The Bribery Act 2010, UK Ministry of Justice). As well as senior management commitment to implementing a culture of integrity, transparency and compliance; the adoption of internal codes of conduct; implementation of whistleblowing systems; mapping risks and implementing internal controls and audits; and the training of staff on corruption risks (French Anti-corruption Agency (AFA) Guidelines 2016).

Furthermore, AI-based insurance might shape the subjectivity of policy holders in un-intended ways (e.g., Kellogg, Valentine and Christin, 2020). For instance, policy holders might change their behaviors because of the constant perception of surveillance (when the policy is perceived to police behaviors to trigger compliance with policy's expectations); constant visibility and loss of privacy (when the policy enables mechanisms of social comparison); and perception of unfairness (when the policy holder cannot make sense of the decisions taken by the AI powering the policy). Independently, and/or jointly, these mechanisms might trigger the emergence of new "gaming" behaviors and/or new social stratification (Lanzolla, Quy. Pesce, 2020).

Auditability in the insurance sector context:

Current regulation ensures appropriate governance of operations and IT risks and adherence to internal governance standards. In the context of insurance, *robust internal governance* is essential with clarity on roles, responsibilities and accountabilities including operational risk management obligations that insurers adhere to in their own best interest.

In addition, AI audits should be put in place for the insurance sector. Such audits should be both at an internal and external level by independent authorities placing emphasis on AI systems that affect (directly and indirectly) fundamental rights. At an internal level, strengthened quality assurance processes for model oversight and controls should be considered so as to demonstrate that internal governance systems are robust and uncompromised enough to address challenges resulting from the use of non-linear AI models. More specifically, before the launch of insurance-specific algorithm-based systems, thorough assessment should be conducted, including a detailed impact and risk assessment. The performance of these systems must be continuously monitored and assessed, throughout the product's lifecycle, by the insurance company and relevant external independent authorities. Insurance companies must put in place the necessary internal governance mechanisms and measures to ensure legal compliance of the AI-systems as well as necessary AI audits.

Trade-offs in the insurance sector context:

A number of trade-offs can be identified in the insurance context, as it can be seen in the Table 1.

Issue	Benefit	Cost
~~Discrimination~~ Risk profiling	Accuracy of Risk Classification	Equal Treatment
~~Intrusiveness~~ Privacy	Risk Reduction	Intrusiveness
~~Secondary use~~ Accuracy	Value of Data	Contextual Integrity
~~Individualisation~~ Pricing	Individual Pricing	Affordability
Solidarity (Social Insurance)	Individualization	Equity
Risk Pooling (Private Insurance)	Individualization	Value of Insurance

(Adopted from Geneva Association)

Redress in the insurance sector context:

Redress processes in the insurance sector are critical. Insurers should develop such processes that will essentially act as a mechanism to compensate for any harm caused by AI (AI4People, 2018). Such processes will foster public trust in AI. Reliable

redress mechanisms for harms inflicted, costs incurred, or other grievances caused by the technology should be established, including a clear and comprehensive allocation of accountability to humans and/or insurance companies. Aligned with the AI4People, 2018 report, the aerospace industry, could be seen as an example, as it has a proven system of handling unwanted consequences thoroughly and seriously. The design process of effective remedies should involve prompt and adequate reimbursement and redress for any harm suffered by the development, deployment or use of AI systems, and may include measures under civil, administrative, or, where appropriate, criminal law (Council of Europe, 2019). Therefore, algorithm-systems and tools in insurance should go through a thorough assessment before their launch, including a detailed impact and risk assessment. In addition, the performance of these system should be regularly monitored and assessed throughout the lifecycle of the insurance product/service/system, by the insurance companies as well as relevant, specialised public authorities.

Establishing accountability

"A 2018 study by the FCA found that several insurance companies were unable to name a dedicated member of staff who had ownership over their pricing strategy (which could include how AI-led risk assessments influence premiums)." Centre for Data Ethics and Innovation on AI and Personnel Insurance (CDEI)

RECOMMENDATIONS - Requirements 1, 4 & 7:

1. **AI systems should empower individuals to make informed decisions**. At the same time, proper oversight mechanisms need to be ensured in the insurance sector. Such mechanisms can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches.

2. **Internal guidelines and policies**: Insurance companies need to build trust through transparency. At an internal level they need to develop and implement guidelines and policies that will ensure transparency and safeguard the interests of employees, stakeholders, customers and the wider community. This way insurers can mitigate the risk, ensuring a consistent approach to the transparency and explainability of algorithmic outcomes. Guidelines should help to clarify how the benefits and risks of using AI should be assessed on a case-by-case basis. Actuaries, risk managers, data scientists and data protection officers should closely cooperate in the development and implementation of such guidelines and policies.

3. **Ethical AI audits for algorithmic development and deployment** in the insurance sector could ensure moral and ethical compliance while at the same time identify potential biases or flaws. Such algorithm audits could be external (top-down) as well as internal (bottom-up, company-driven) auditing mechanisms that are performed at regular intervals.

4. **Accountability via governance frameworks** (HLEG recommendation). Robust internal governance with clarity on roles, responsibilities and accountabilities is covered by operational risk management obligations that insurers adhere to in their own best interest. Strengthened quality assurance for model oversight and controls should be considered to demonstrate that internal governance systems are robust and uncompromised enough to address challenges resulting from use of non-linear AI models.

5. **An AI system should be designed to be fully answerable and auditable**. In order for this to be achieved, it is important to: (a) establish a "continuous chain of responsibility" for all roles involved in the design and implementation lifecycle of the project, and to (b) implement a continuous "activity monitoring" so as to allow for oversight and review throughout the entire project lifecycle.

6. **Insurers should develop appropriate and effective redress processes** that will involve prompt and adequate reimbursement and redress for any harm suffered by the development, deployment or use of AI systems, in direct alignment with the relevant laws. Thorough assessments (impact, risk assessments, human rights assessments, etc.) of AI insurance systems and tools should be conducted under distinct time periods (pre-launch, post-launch) and throughout the lifecycle of the insurance product/service/system, by the insurance companies as well as relevant, specialised public authorities.

The insurance sector cannot be considered in isolation of other industries it enables. Changes in those industries will inevitably inform insurers' ability to go by the above recommendations. Whether in human agency, explainability or accountability, insurers depend heavily on the modus operandi of other sectors. The other way around, they may be used as a 'canary in the coalmine', able to tell what is insurable from what is not, which is more or less a way to tell what is socially correct and what is not, what is worth putting on the market and what is not.

This is how "Governance Innovation" describes possible benefits accruing from a cross-sectoral approach: "Businesses that appear to be distinctly different from one another when viewed from a physical space perspective may share many cross-sectorial commonalities when seen from a cyberspace operations perspective. For example, we

believe that cross-sectorial goals as well as guidelines and standards can be defined for areas such as data management (privacy, cyber-security), ID infrastructure construction, AI quality assessment and continuous data collection method. Because it is currently difficult to ensure the predictability and explainability of the performance outputs of complex systems centered on software—such as AI that perform machine learning—using conventional rule-based software authentication methods, in order to implement these systems in our societies we may establish more flexible evaluation criteria and construct technological foundations for evaluations that allow for AI and similar technologies to be introduced into the market to a certain level and allow performance improvements to be made through updates. Furthermore, as described in 5.3.2, we can deepen discussion on the behavior of complex systems including AI regarding incentivized regimes for investigations performed by businesses, criminal deferred prosecution agreement regimes, and the establishment of insurance mechanisms that cover potential dangers to society." (page 67 of the Japan METIs report 2019).

3.2 Req. 2: Technical robustness & Safety

REQUIREMENT 2: Technical robustness and safety - ***Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility***

Potential forms of security-related attacks in the insurance sector include the following:

- **Potential forms of security related attacks:**
    - o "*While insurers may be tempted to store this data, perhaps in the expectation they will be able to put it to use in future, doing so raises several ethical concerns. One is the threat to people's privacy, especially where datasets are at risk of a cyber breach. Another relates to fair compensation [for selling this data].*"(Centre for Data Ethics and Innovation on AI and Personnel Insurance)

Examples of Cybersecurity Incidents in the Insurance Sector

The insurance sector has experienced a variety of cybersecurity incidents in past few years, including well-publicised data breaches at several U.S. health insurers. According to the International Association of Insurance Supervisors (IASIS) report (2016), such cybersecurity incidents include the following:

Data breach: In 2015 in the United States, Anthem Blue Cross Blue Shield and Premera Blue Cross (the largest health benefits company by membership in the US) discovered data breaches that began a year earlier. The breached data included member

and applicant names, dates of birth, Social Security numbers, bank account information, claims data, member identification numbers, and some clinical data. The sophisticated cyberattack that lasted for nearly one year before it was discovered, potentially exposed the personal information of up to 91 million policyholders – as many as a quarter of the people in the United States[10]. The insurers had to react swiftly to mitigate reputational damage and to minimise litigation costs. They paid $260m for security improvements and remediation, and $117m in June 2017 to settle lawsuits from customers potentially affected[11]. In addition, during 2019 Premera achieved HITRUST-certification that demonstrates the company's ability to identify risks, protect data, detect attacks, and respond to security incidents.

***Distributed denial-of-service (DDoS) attack*:** DDoS attack threats by a group of cyber extortionists known as the "DD4BC" that had been targeting a range of firms including financial institutions in Europe, Australia, Canada, and the United States in order to extort money from them. DD4BC demanded ransoms, from specific targets, to be paid in crypto-currency (bitcoin), in order to cancel the launch DDoS attacks. Two German insurance groups experienced this type of attack in mid-2015, receiving threats of a DDoS-attack on company web servers. The insurers refused, as they assessed the extortionists would have caused only minor damage in those instances, but these incidents could have been far more serious if the attacks had concentrated on more critical systems.

Cyber-attack: In the Netherlands, an insurer was recently subject to the so-called "CEO hack," a specific form of phishing cyber-attack. Pretending to be the CEO of a major and well-known commercial customer of the insurer, the criminals tried to persuade employees of the insurer to transfer money into a certain account. The criminals had apparently researched certain operational details of the insurer.

- **Potential forms of attack**. There are several categorization of threats, and they all mostly relate to data stored on servers and information exchanged in communications; unintended human actions facilitating cyber-attacks, outages or malfunctioning, physical attack, or attacks to sensors or AI systems or update procedures.

1. For ICT in general, the EU Cybersecurity Act establishes a certification framework, and allows the "creation of tailored and risk-based EU certification schemes". https://ec.europa.eu/digital-single-market/en/eu-cybersecurity-certification-framework

10 Anthem, Inc., Statement Regarding Cyber Attack Against Anthem, via https://www.anthem.com/health-insurance/about-us/pressreleasedetails/WI/2015/1813/statement-regarding-cyber-attack-against-anthem; Premera Blue Cross, Premera Targeted by Cyberattack (17 March 2015), via https://www.premera.com/wa/visitor/about-the-cyberattack/?WT.z_redirect=www.premera.com/cyberattack/

11 https://healthitsecurity.com/news/premera-reaches-proposed-74m-settlement-over-2014-breach-of-11m

· **Applicability of relevant Insurance Core Principles to Cybersecurity (ICPs)**: despite the fact that ICPs do not specifically address cyber risk and cyber resilience they provide a general basis for managing cyber risks and information exchange. Relevant ICPs are: ICPs 7, 8, 9, 19, 21 (cyber risks), ICPs 3, 25, 26[12].

Protecting Customer Data through cybersecurity

Data is creating a new horizon of opportunities for insurance companies. In order to realise its full potential, insurance firms must respect and protect their customers and their data. Therefore, insurance as a business of trust, needs to take all the necessary measures to protect data and demonstrate to customers the data privacy commitments, going beyond mere compliance with privacy and data breach laws.

Focusing upon the protection of customer data through cyber security, Zurich Insurance Group has established the Cyber Fusion Center as an internal cyber threat intelligence group. The Center aims to protect customers' data by combining cyber-threat intelligence, response, forensics and vulnerability management teams.

As cyber-attacks increase both in frequency as well as in severity, companies can make themselves more resilient by strengthening their ***cyber-risk strategies and practices*** at all levels of the institution and with respect to relevant third-party arrangements. However, the systemic nature of continuously evolving cyber threats, necessitates the *need for collective action*. As for insurers, cybersecurity incidents can harm not only the ability to conduct business, compromise the protection of commercial and personal data, and undermine confidence in the insurance sector (causing reputational damage affecting the confidence of consumers, policy holders, investors, rating agencies and business partners)[13]. One example towards such collective action is the **Centre of Cybersecurity[14] of the World Economic Forum**, which brings together experts from around the globe aiming to address systemic cybersecurity challenges and improve digital trust to safeguard innovation, protecting institutions, businesses and individuals.

RECOMMENDATIONS - Requirement 2:

Insurers need to be alert to cybersecurity risks, to put relevant safeguards in place and should utilise the certification framework set out in the EU Cybersecurity Act.

12 ICP 7 (Corporate Governance), ICP 8 (Risk Management and Internal Controls), ICP 9 (Supervisory Review and Reporting), ICP 19 (Conduct of Business), ICP 21 (Countering Fraud in Insurance), ICP 3 (Information Exchange and Confidentiality Requirements), ICP 25 (Supervisory Cooperation and Coordination), ICP 26 (Cross-border Cooperation and Coordination on Crisis Management) (International Association of Insurance Supervisors- IASIS, Issues paper on cyber risk to the insurance sector (Aug. 2016)).

13 International Association of Insurance Supervisors (IASIS), Issues paper on cyber risk to the insurance sector (Aug. 2016).

14 Centre of Cybersecurity of the World Economic Forum https://www.weforum.org/platforms/shaping-the-future-of-cybersecurity-and-digital-trust

3.3 Req. 3: Privacy & Data Governance

REQUIREMENT 3: Privacy and data governance
Including respect for privacy, quality and integrity of data, and access to data

Data and data processing constitute a core aspect of the insurance sector. Digitisation has enabled, insurance companies to enrich their traditional datasets (e.g. demographic data, exposure or behavioral data, etc.), with new types of data such as Internet of Things (IoT) data, online data, bank/credit data, etc., and to perform highly advanced analytical processing. Across the insurance value chain, the data used (EIOPA, 2019) can: (a) include personal data (e.g. medical record) and non-personal data (e.g. hazard data), (b) they can be structured (e.g. IoT data, survey) or unstructured (e.g. pictures or emails) and (c) can be obtained from internal sources (e.g. consumer provided data directly to the firm) and from external sources (e.g. public databases, private data vendors, etc.) (see Table 2).

Table 2: Traditional and New data sources in the insurance sector

Traditional data sources	New data sources enabled by digitalisation
Medical data (e.g. medical history, medical condition, condition of family members)	IoT data (e.g. driving behaviour (car telematics), physical activity and medical condition (wearables)
Demographic data (e.g. age, gender, civil and family status, profession, address)	Online media data (e.g. web searches, online purchases, social media activities, job career information)
Exposure data (e.g. type of car, value of contents inside the car)	Insurance firms' own digital data (e.g. interaction with insurance firms (call centre data, users' digital account information, digital claim reports, online behaviour while logging in to insurance firms' websites or using insurance firms' app)
Behavioural data (except IoT data) (e.g. Smoking, drinking behaviour, distance driven in a year)	Geocoding data (i.e. latitude and longitude coordinates of a physical address)
Loss data (e.g. claim reports from car accidents, liability cases)	Genetics data (e.g. results of predictive analysis of a person's genes and chromosomes)
Population data (e.g. mortality rates, morbidity rates, car accidents)	Bank account / credit card data (e.g. consumer's shopping habits, income and wealth data)
Hazard data (e.g. frequency and severity of natural hazards)	Other digital data (e.g. selfie to estimate biological age of the consumer)
Other traditional data (e.g. credit scoring, claim adjustment reports, information from the auto repair shops)	

(Sources: EIOPA, 2019; Geneva Association, 2020)

The use of demographic data (e.g. age, gender, occupation, etc.) is a widely adopted practice among insurance firms. Demographic data are directly provided by the consumers at the pre-contractual stage and are subsequently complemented with external data sources such as national statistics offices or third-party data vendors (e.g. geo-spatial socioeconomic demographic classifications, such as purchasing power, family types, population density etc.). (EIOPA, 2019). Based on a recent study by EIOPA (2019), health and motor insurance firms collaborate with third-party data vendors in order to acquire anonymised third-party data (at a postal code and granular level) that are used in technical models for pricing and underwriting purposes. Insurance firms also collect gender data, which they are not allowed to use for pricing and underwriting purposes following the 2011 ruling of the European Court of Justice against the pricing differentiation on the grounds of sex (European Court of Justice, 2011)[15].

Digitisation enables new types of data (e.g. IoT data). According to EIOPA (2019), the usage of such data is expected to increase in the next three years in the insurance sector, primarily in the motor insurance sector (e.g. data collected via black boxes installed in cars or mobile phone apps including speeding data, miles driven, road types, harsh braking, etc.) as well as in other sectors such as the health insurance sector (e.g. data collected via wearables or mobile phone apps). This information often complemented with external data sources has and will facilitate pricing and underwriting (e.g. Pay-As-You-Drive (PAYD) or Pay- How-You-Drive (PHYD) policies (Usage-Based Insurance)).

Emerging sources of data in insurance and ethical concerns

New sources of data are entering the insurance sector, such as data from wearables and telematic devices. In this emerging context, according to the Centre for Data Ethics and Innovation (CDEI), insurers may find themselves collecting more information about their customers than is necessary to deliver their core services. *However, should insurers store this data with the expectation to put it to use in the future?*

This question raises several ethical concerns. One is the threat is people's privacy, especially where datasets are at risk of a cyber breach. Another concerns fair compensation, if for example customer data is subsequently sold onto third parties. This raises aspects of adequate reimbursement of customers for the value that they have created for the insurance company.

Furthermore, the collection of additional data may also increase the probability of algorithmic bias during the training phase. In order to reduce these harms, the industry could prepare necessary standards, for example **data storage standards**, that could be **jointly developed with relevant Insurance Associations or Standards Institutes**,

15 Case C-236/09, The European Court of Justice, 1 March 2011, http:// curia.europa.eu/

at a national and regional level, aiming to discourage insurance companies from storing data that is not directly linked to their mission. Such standards, according to CDEI, could include an expectation for insurers to regularly review their datasets so as to determine whether they are material to the ir core business practices, and if not to eliminate them from company records.

- **Existing EU rules on privacy and data:**
  - o Controllers must, at the time when personal data is obtained, provide the data subjects with information necessary to ensure transparent processing about the existence of automated decision-making. The GDPR recognizes the overarching principle of fair treatment of consumers in relation to personal data processing, and enables consumers to demand the removal of their data from the insurer's databases ("right to be forgotten", Art.17(2) GDPR). In addition to the strong obligations for organisations, GDPR introduces several accountability tools to data protection rights such as the principles of data protection by design and by default and new provisions for company certification and industry-wide code of conduct schemes (General Data Protection Regulation (GDPR) and Data Protection Law Enforcement Directive).

Privacy and data protection in the insurance sector:

The level of governance of data and model accuracy depends on the business purpose (e.g. for medical related services the scrutiny level must be much higher than for marketing purposes). In general, AI insurance applications in customer engagement may have lower significance than applications that are used to determine payouts to policyholders. Applications in underwriting/pricing may exhibit the highest levels of significance, in particular if they could lead to the exclusion of customers and impact their privacy. For example, behavior change schemes could pose a threat to the autonomy of policyholders, with insurers gaining the power to influence their lives in multiple ways, from where they live to how they drive to how often they exercise. While one could argue that signing up to behavior change schemes is a choice, it would be relatively simple for insurers to turn a voluntary scheme into a mandatory one. Refusing to participate in such schemes may also signal to insurers that customers are high-risk, since low-risk individuals would have every incentive to be monitored.

In addition, other uses of AI that automate specific tasks but do not change the logic of decision-making in any way (such as extracting relevant information from documents in an automated way via OCR and NLP) are likely to exhibit low significance of impact. In contrast, any use of AI that changes the logic of decision-making (i.e. that applies a new model to existing data) may exhibit higher significance. The highest level

of significance of impact may be in uses that change the logic of decision-making based on new data sources (see Figure 1).

Ethical Review Boards

Acknowledging the importance of the ethical use of algorithms and data, some insurers are increasing their level of disclosure in relation to these aspects. For example, the Zurich Insurance Group, launched a Data Commitment, and **UK Aviva insurance company**, launched a **Customer Data Charter**[16] that aim to set out what happens to the information they collect on customers, and the corporate rules around how it is shared – aiming to protect personal data. Other insurance firms have established expert panels that focus upon their corporate policy.

For example, **AXA's Data Privacy Advisory Panel**[17] is an internal team that is composed by data and privacy experts, business and marketing executives, legal practitioners and corporate responsibility officers as well as academics and independent advisors. The panel meets bi-annually in order to consider AXA's use of data and algorithms, as well as the firm's actions and commitments (e.g. data privacy commitments, including aspects that relate to the international exchange of data, among others).

This company-centric, *bottom-up approach* aims to foster ethical practices and the ethical use of data and algorithmic systems. This approach entails the development of ethical advisory boards or panels, integrating a wide range of company-members from different departments. Such panels tend to have an advisory, review role including for example the review of insurance applications, the review of potential violations of the internal ethical code, as well as the review of new software that contains AI elements and the development of actual measures, among others. They can be linked to an industry-wide Code of Conduct scheme, which is encouraged under Article 40 of the GDPR as well as company certification schemes.

RECOMMENDATIONS - Requirement 3:

1. Compliance with data protection standards. Insurers should ensure compliance with the General Data Protection Act and the Data Protection Act, among other legislation, placing emphasis on personal data storage and being clear on the legal basis for processing data (CDEI, 2019).

2. Disciplined internal oversight of the data science pipelines and strong business data ownership and accountability can support the accuracy of data and mitigate privacy and data risks.

16 Zurich's data commitment, https://www.zurich.com/media/magazine/2019/earning-trust-to-unlock-the-power-of-data; Aviva's customer data charter, https://www.aviva.co.uk/services/about-our-business/about-us/customer-data-charter/
17 AXA's data privacy advisory panel was set up during July 2015, https://www.axa.com/en/about-us/data-privacy

3. The formation of ethical advisory/review boards can be seen as a bottom-up approach to establishing ethical AI practices and Trustworthy AI. The members of these boards should have sufficient diversity of expertise such as technical, ethical, legal and experts on the subject matter (domain of application). Insurance firms aligned with the HLEG recommendations can adapt their corporate responsibility charter, Key Performance Indicators (KPIs), codes of conduct or internal policy documents to add the striving towards Trustworthy AI.

4. The **fair, ethical and transparent use of data is a priority**. Data analytics governance frameworks can be utilized in order to create trust and ground the use of data in common ethical principles (EIOPA, 2020).

3.4 Req. 5: Diversity, non-discrimination and fairness

REQUIREMENT 5: Diversity, non-discrimination and fairness

Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

Fairness in a key and highly important principle for responsible AI, that is associated with distinct values - such as freedom, dignity, autonomy, privacy, non-discrimination, accessibility, equality and diversity, among others. These values often need to be interpreted in context, including the cultural context, which makes it impossible to provide a universal standard of fairness (GA, 2020). At a general level, distinct dimensions of fairness can be distinguished such as procedural and substantive (European Commission 2019):

- ***fair process*** – consumers fair treatment across the whole process; where a core aspect of fair treatment is the ability of customers to challenge and seek effective redress against decisions affecting them (European Commission 2019). In the context of the insurance sector there are market conduct requirements to ensure fair treatment of customers irrespective of the technology used (such as Insurance Core Principle 19 (ICP19) of the International Association of Insurance[18]) (GA, 2020).
- ***fair decision making*** – AI-driven decision making should be fair so that it does not unfairly discriminate and disadvantage individuals or groups of individuals (European Commission 2019, 2020; as well as other ethical AI guidelines for e.g. OECD 2019). Thus, non-discrimination and avoidance of unfair bias are

18 The Insurance Core Principles (ICPs) developed by the International Association of Insurance Supervisors (IAIS) provide a globally accepted framework of principles, standards, and guidance for the regulation and supervision of the insurance sector. In the context of fairness, the IAIS-Insurance Core Principle 19 (ICP19) states that 'The supervisor requires that insurers and intermediaries, in their conduct of insurance business, treat customers fairly, both before a contract is entered into and through to the point at which all obligations under a contract have been satisfied.'

core aspects of fairness. In addition, the equal and just distribution of both benefits and costs, constitute another key aspect of fair decision making (fairness of outcome) (European Commission 2019).

At an AI system level, diverse dimensions of fairness can be distinguished along its lifecycle, so that it meets a minimum level of discriminatory non-harm. These include: ***data fairness*** – usage of fair and equitable datasets only, ***design fairness*** – inclusion of reasonable feature, processes, and analytical structures in the model architecture, ***outcome fairness*** – prevent the system from having any discriminatory impact, and implementation fairness – implement the systems in an unbiased way.

In the insurance context, this requirement can be associated with a low to medium risk depending on the insurance use-case. For example, for some use cases such as Life & Health, hyper-personalization, pricing and underwriting, if fairness and discrimination aren't properly assessed at every step of the value of chain.

Avoiding bias (i.e. unwanted discrimination) is a technical question that can be resolved with sound methodology for AI and algorithms. Methods to avoid unwanted bias in AI should be developed and applied in line with scientific progress. For example, in order to limit adverse impacts of enhanced pricing algorithms on customers, such as bias and implications for affordability, insurers can implement a comprehensive post-monitoring system. Models are scrutinized by diverse teams from different functions, and system output is tested using different tests to validate that the models are in line with expectations of the modelling team, business partners and regulators.

Where regulatory approval of rates is required, regulators can be provided with dedicated tools to track the performance of the system. Particular attention should be given to potentially high-risk applications (e.g. robo-advice for asset allocation or long-term financial products), which could be considered for ex-ante regulatory approval.

Fairness issues can also be addressed by a broader ethical discussion and adherence to a proper conduct regime and transparency on underlying value-based decisions. The existing conduct regime for the financial services sector already provides a robust principle-based ethical framework (such as actuarial ethics). However, interpretation to resolve ambiguities is required.

To monitor and mitigate bias requires the quantification of fairness. While academic research on appropriate fairness metrics is still evolving and should be encouraged[19], insurers need to identify context-specific fairness definitions for each use of AI.

19 Singapore Monetary Authority Services (MAS) launched in May 2020 a framework for financial institutions to promote the responsible adoption of Artificial Intelligence and Data Analytics (AIDA). It will commence with the development of fairness metrics in credit risk scoring and customer marketing. https://www.mas.gov.sg/news/media-releases/2020/fairness-metrics-to-aid-responsible-ai-adoption-in-financial-services

Practical frameworks for assessing fairness require a suitable balance to be found between four (potentially competing) objectives[20]:

- business model requirements—sustainable and profitable product provision, which is in the long-term interests of customers, subject to market conditions and competitive dynamics;
- ensuring consumer value—meeting the fundamental demand for the product;
- consumer protection—ensuring good outcomes and avoiding the exploitation of vulnerable customers;
- consumer choice—considering the alternatives that customers may or may not have, hence avoiding the exploitation of customers with limited alternative

Roles and responsibilities with respect to monitoring and mitigating bias should be clearly defined. As bias can enter decision-making at various stages, it is important to raise awareness at different management levels through appropriate educational and training programs.

- "ICO[21] advises organisations on how they can adhere to the Data Protection Act and the GDPR, among other legislation. It cuts across every sector and affects the majority of organisations, including within the insurance industry. Recent and relevant ICO initiatives include Project ExplAIn, which will assist organisations as they attempt to explain the results of AI decision-making; and the development of an Auditing Framework for AI, that will guide the regulator's efforts in examining algorithms for fairness." (Centre for Data Ethics and Innovation on AI and Personnel Insurance).

Avoidance of unfair bias in the insurance sector:

Centre for Data Ethics and Innovation on AI and Personnel Insurance – "Insurers are prohibited by law from basing pricing and claims decisions on certain protected characteristics, including sex and ethnicity. However, other data points could feasibly act as proxies for these traits, for example with postcodes signalling ethnicity or occupation categories signalling gender. This means that AI systems can still be trained on datasets that reflect historic discrimination, which would lead those systems to repeat and entrench biased decision-making. A Propublica investigation in the US found that people in minority neighbourhoods on average paid higher car insurance premiums than residents of majority-white neighbourhoods, despite having similar accident costs. While the journalists could not confirm the cause of these differences, they suggest biased algorithms may be to blame.

20 Oxera's 'Fair ground: a practical framework for assessing fairness', Oxera Agenda March 2019.
21 The Information Commissioner's Office is the UK's principal regulator for upholding information rights in the public interest.

Online car insurance and Discrimination Claims

Online car insurance companies use predetermined algorithms to assess the risk of a user filing a claim against their policy. In 2018, a public backlash[22] started building against large global firms like Admiral, Marks & Spencer, Bell, Elephant and Diamond, as it was found that insurance quotes for drivers with traditional English names, like 'John', were far lower than quotes of the same for drivers with non-English names, such as 'Mohammed' for example, for identical insurance details.

The insurance company Admiral was under scrutiny as it was found that when applying for quotes via the price comparison website GoCompare, the same insurance for a 2007 Ford Focus in Leicester was priced at £1,333 for 'John Smith' and £2,252 for 'Mohammed Ali'.

This was further validated, after obtaining sixty quotes ranging across ten different cities through a number of price comparison websites including GoCompare. According to the newspaper that obtained the quotes, in all cases the difference was often hundreds of pounds. Following these concerns and complaints, the UK's Financial Conduct Authority (FCA) put under close scrutiny the pricing practices of UK insurers and intermediaries, placing emphasis in dual pricing and discrimination[23].

FCA's Interim Report of the general insurance pricing practices (FCA, 2019) identified a number of concerns. In relation to differential pricing, FCA's research found evidence of differential pricing between, long standing customers who continually renewed with their existing provider (tended to pay higher – a so called "loyalty premium") and new customers. Furthermore, FCA, set out concerns about how pricing in these markets leads to consumers who do not switch or negotiate with their provider paying high prices for their insurance.

In relation to discrimination against protected characteristics FCA (2018) report, found no evidence of firms engaging in direct discrimination, it voiced concerns about the potential use of data based on race/ethnicity within firms pricing models. According to FCA, firms pricing models were found to be using datasets, including third party datasets (that were used without always undertaking the necessary due diligence so as to ensure that data exclude factors that may lead to discrimination based on protected characteristics) that could "contain factors that could implicitly or potentially explicitly relate to race or ethnicity[24]."

22 Money Saving Expert (2018). Accessed via: link.
23 "Dual Pricing in Insurance – the beginning of the end?", April 10, 208, Retrieved from: link
24 "When firms were asked how they gained assurance that the third-party data they used in pricing did not discriminate against certain customers based on any of the protected characteristics under the Equality Act 2010. Many firms could not provide this assurance without first contacting the third-party provider. Further, some firms responded that they relied on the third-party provider to comply with the law and undertook no specific due diligence of their own to ensure that the data were appropriate to use." (FCA, 2018, p. 15).

Like any organisation using algorithms to make significant decisions, insurers must be mindful of the risks of bias in their AI systems and take steps to mitigate unwarranted discrimination. However, there may be some instances where using proxy data may be justified. For example, while car engine size may be a proxy for sex, it is also a material factor in determining damage costs, giving insurers more cause to collect and process information related to it. Another complication is that insurers often lack the data to identify where proxies exist. Proxies can in theory be located by checking for correlations between different data points and the protected characteristic in question (e.g. between the colour of a car and ethnicity). Yet insurers are reluctant to collect this sensitive information for fear of customer believing the data will be used to directly discriminate against them." As a general principle, AI systems should be trained to guarantee all human beings fair and equal treatment with no distinctions among age, gender, race, etc..

Use cases for Requirement 5:

Flood Insurance: A UK initiative that balances fairness, regulation and accuracy

In the UK, flooding is recognized as a common natural disaster and flood damage coverage is available to residential customers and small businesses as part of the standard terms of property insurance. During the past few years advances in the use of Geographic Information Systems, remote sensing and simulation modelling and detailed hydrological flood models, as well as the ability to access new high-quality data, have significantly improved the ability of insurers to assess flood risks. This resulted in a highly segmented home insurance market, with an informal cross-subsidy between low- and high-risk homes[25].

However, the increasing affordability problems for the individuals at high risk, lead to a debate about fairness. So, should those at high risk pay a premium to match, even if unaffordable, or should they be supported by cross-subsidy from the rest of the population?

Following the serious UK flooding in 2000 (affecting 10,000 properties in 700 locations and caused £1 billion of damage[26]), the UK government and the Association of British Insurers (ABI) drew up a Statement of Principles (incorporated in the Gentlemen's Agreement[27]). This agreement entailed that ABI members would continue to offer insurance at existing rates to properties at high risk of flooding, if the government continued to invest in flood defences. After the expiration of the Statement of Principles (2013), the fairness debate spurred again leading to the emergence of the Flood Re partnership, as a not-for profit fund owned and managed by the insurance industry.

25 Cullen, M. (2015) The ABI view: Sharing risk or smoothing bad luck, Insurance Times, 19 October 2015. Available at: https://www.insurancetimes.co.uk/the-abi-view-sharing-risk-or-smoothing-bad-luck/1415939.article

26 Environment Agency (2001), Lessons learned Autumn 2000 floods, March 2001. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/292917/geho0301bmxo-e-e.pdf

27 The unwritten Gentlemen's Agreement between the UK Government and private sector insurers, facilitated the private sector flood insurance operations in the UK since the early 1960s (Huber M. (2004), The Breakdown of a Gentlemen's Agreement, Discussion Paper No. 18, London School of Economics and Political Science, ESRC Centre for Analysis of Risk and Regulation, 2004. Available at: http://eprints.lse.ac.uk/36049/1/ Disspaper18.pdf.)

Flood Re[28] is a joint initiative between the Government and insurance industry that aims to make the flood cover part of household insurance policies more affordable. This flood re-insurance company essentially provides flood insurance coverage to the vast majority of households (domestic properties) and small businesses deemed at significant risk of flooding[29] and where no public plans are in place to defend the property (The Government and the Association of British Insurers, 2008[30]). The scheme that enables insurance companies to insure themselves against losses because of flooding, became operational in 2016 with a 25-year lifetime, becoming the first scheme of its kind anywhere in the world.

RECOMMENDATIONS - Requirement 5:

1. **Fairness via the mitigation of discrimination and bias in the insurance sector is challenging, *but joint efforts, initiatives and partnerships between stakeholders***, like Public Private Partnerships (e.g. Flood Re), etc. can increase our social resilience, solidarity (insurability) and social responsibility. Monitoring is an important parameter that facilitates the discrimination and bias mitigation processes. The Dutch Insurance Association has developed a "solidarity monitor" to assess the spread of insurance premiums and individual insurability across time.

2. **Insurance companies can eliminate biases with testing and monitoring (auditing mechanisms)**. Insurers should test and continuously monitor AI systems for unwanted consequences, such as unfair biases, in which case they should remove data sets that create or preserve them, and take human-led corrective action. In addition, AI models should be scrutinized by diverse teams from different functions, and system output should be tested using different tests to validate that the models are in line with expectations of the modelling team, business partners and regulators. Where regulatory approval of rates is required, regulators can be provided with dedicated tools to track the performance of the system. Internal roles and responsibilities with respect to monitoring and mitigating bias should be clearly defined.

28 Flood Re, https://www.floodre.co.uk/

29 This is generally defined as more than a 1.3% or 1 in 75 annual probability of flooding – Based on The Government and the Association of British Insurers 2008, "Revised statement of principles on the provision of flood insurance", July 2008. Available at: https://www.abi.org.uk/globalassets/sitecore/files/documents/publications/public/migrated/flooding/statement-of-principles-england.pdf

30 The Government and the Association of British Insurers (ABI) 2008, "Revised statement of principles on the provision of flood insurance", July 2008. Available at: https://www.abi.org.uk/globalassets/sitecore/files/documents/publications/public/migrated/flooding/statement-of-principles-england.pdf

3. **Particular attention should be given to potentially high-risk applications** (e.g. robo-advice for asset allocation or long-term financial products), which could be considered for ex-ante regulatory approval.

4. **Educational and training programs:** As bias can enter decision-making at various stages, it is important to raise awareness at different management levels through appropriate educational and training programs

5. In addition to developing auditing mechanisms for AI systems, a **solidarity mechanism** to deal with severe risks in AI intensive sectors should be developed (AI4People, 2018). Those risks could be mitigated by multi- stakeholder mechanisms upstream. Pre-digital experience indicates that, in some cases, it may take a couple of decades before society catches up with technology by way of rebalancing rights and protection adequately to restore trust. The earlier that users and governments become involved – as made possible by ICT – the shorter this lag will be. (AI4People, 2018).

3.5 Req. 6: Societal and environmental wellbeing

REQUIREMENT 6: Societal and environmental wellbeing
Including sustainability and environmental friendliness, social impact, society and democracy

AI systems should be used to benefit all human beings, current and future generations as well as the natural life. Consistent with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be considered as stakeholders throughout the AI system's life-cycle (AI HLEG, 2019). AI systems should be used to enhance positive social change and encourage sustainability and environmental responsibility of such systems.

Sustainable and environmentally friendly AI in the insurance sector: The environmental wellbeing issue is an emergent topic for the insurance sector, for example the EU agenda on digitalization for a transition to a more sustainable future is using AI to predict and mitigate climate change related risks. As such it is important to encourage measures securing the **sustainability and environmental friendliness** of AI systems in the insurance sector. For example, selecting a low energy consumption method and increasing the environmental sustainability of an AI system in insurance.

Social impact of AI in the insurance sector: In addition, the **social impact** of AI systems in the insurance sector should be enhanced. For example, AI applications can

help to extend insurance cover to new and previously uninsured or underinsured customer segments or to expand the range of risks for which insurance cover is available (Geneva Association, 2020). As such, AI can facilitate the expansion of the scope of risk pooling, which lies at the core of the economic and societal role of insurance (see Table 3), according to recent study by the Geneva Association. The social impacts of AI systems at an internal (i.e. impact on workforce) and external level (i.e. customers) should be encouraged and constantly monitored.

Table 3: Socio-economic benefits of AI for Insurance

Expand the scope of risk pooling
Extend insurance over the new and previously uninsured customer segments by facilitating access to personalised products (e.g., life insurance for individuals with pre-existing conditions)
Expand the range of risks for which insurance cover is available through improved risk insights (e.g., cyber risks)
Reduce the cost of risk pooling
More cost-efficient insurance through the automation of specific tasks, better risk assessments and reduction of moral hazard and adverse selection
Mitigate and prevent risks
Novel risk insights that help mitigate and prevent risks
Early warning systems that enable the reduction of losses

(Source: Geneva Association, 2020)

Society and democracy: The effects (direct and indirect) of AI systems on society at large as well as democracy should be considered and assessed. Special emphasis should be given to electoral contexts (e.g. amplification of fake news, etc.) and cases where AI systems in the insurance sector may have a negative impact. Dedicated actions and measures should be taken in order to minimize the potential harm.

Automated Decision-Making in Insurance: Examples from Europe

According to the recent Algorithmwatch report[31] there are a number of examples of automated decision-making in Europe. In the context of the Insurance sector the following examples were identified across different EU counties.

Denmark: Car Insurance

Profiling and automated decisions are present in the insurance sector in Denmark. These activities are regulated via data protection laws and overseen by the Danish data protection authority, Data Tilsynet. In the context of car insurance, insurance companies offer rebates if drivers install a box (a type of telematics car insurance, often called black-box car insurance[32]) that measures speed, acceleration, deceleration and g-force and in turn drivers are offered a fixed 25% rebate for installing the box (Spielkamp, 2019). In other cases, car insurers created a mobile app[33] (that is currently unavailable) that included driving instructions and measurements resulting in quarterly, monthly, or even more frequent adjustments to the car insurance premiums.

Finland: Benefit process at the Social Insurance Institution

The Social Insurance Institution of Finland (Kela) is settling benefits under the national social security programmes (such benefits include health insurance, state pensions, student financial aid, housing allowances, and basic social assistance among others). Decision automation in Kela, has relied for decades in traditional automated information systems. However, AI, machine learning, and software robotics are seen as an integral part of their future ICT systems (including chatbots for customer service, automated benefit processing, detection (or prevention) of fraud or misunderstanding, and customer data analytics) (Spielkamp, 2019).

Netherlands: Credit/Risk Scoring

An increasing number of private companies offer credit scoring services that are used by numerous clients including health care insurance providers. The credit scoring companies provide an automated indication regarding the creditworthiness of a potential customer. As such, the clients of the credit scoring services, such as insurance companies, can use additional automated decision-making process to decide whether, for example, a potential customer can have insurance. These credit scoring companies exist alongside an official and independent financial registration office (Central Credit

31 A recent report by the Algorithmwatch – a non-profit organisation promoting algorithmic transparency - in cooperation with Bertelsmann Stiftung, supported by the Open Society Foundations, provides a comprehensive study of the state of automated decision-making in Europe, listing examples of automated decision-making already in use (Spielkamp, 2019).

32Black box car insurance involves the installation and use of a type of telematics equipment – known as a black box – which uses GPS to monitor and set car insurance premiums based on the driving habits of the individual. It is one of the most well-known types of telematics car insurance, in addition to: 'Plug-and-drive' (are devices that also use GPS technology, but instead of installing a black box, that car insurer provides a device that individuals can plugs into their car's charging port directly) and 'smartphone app' (mobile app that once installed it can track the driving habits of the individual). In US, the insurance company- American Automotive Association, is providing a specialised smartphone based tool – AAADrive – in the AAA Mobile App that measures the driving habits of individuals and produces a safe driver score. Qualified insurance customers have the ability to view and receive a score of their driving behavior and earn potential discounts of up to 30% on their auto policy (Avi Ben-Hutta 2019, " Introducing AAADrive", Available at: link).

33 See article "Mobilen giver gode bilister rabat hos Nem Forsikring" (2016). Available at: link.

Registration Office or BKR- Bureau Krediet Registratie), however in many cases the amount of data that they collect far exceeds the amount available at the BKR (Spielkamp, 2019).

Slovenia: Insurance

Algorithmic systems are used in order to facilitate insurance agents in a number of cases. The biggest Slovenian insurance company Triglav, is using algorithmic systems in order to assist its insurance agents in recommending appropriate insurance products to customers, to detect fraud and to assess insurance risks. However, these systems are only used as counsellors and the final decision is taken by their human agents, according to a spokesperson.

Sweden: Automated Home Insurance

The Swedish, home insurance start-up company Hedvig uses technology to automate many insurance processes, such as pricing and filling of insurance claims. The Hedvig app uses a voice input in order to automatically write and send the claim so that the insurance company can automatically process the claim and disburse the payment.

RECOMMENDATIONS - Requirement 6:

1. **AI systems should benefit society and the natural environment, now and in the future**. Therefore, "we need to adopt a culture of AI systems development that is both *socially-good-by-design and environmentally-friendly-by-design*" (Ziouvelou and McGroarty, 2021). Towards that end, we should use financial incentives to lead development and use of AI technologies towards socially preferable (not merely acceptable) and environmentally friendly (not merely sustainable but favorable to the environment) outcomes (Floridi et al., 2018). This is important both at an overall EU level and at individual national levels. This will entail putting in place structures and methods for assessing the socially goodness and environmentally friendliness of AI systems and projects.

4. Recommendations for the insurance sector

4.1 Recommendations for each of the 7 Key requirements

The table below provides an overview of the key requirements for a Trustworthy AI and the associated risk levels from the industrial perspective.

Table 4: Industry Perspective on Trustworthy AI: Key Requirements, Associated Risks and Mitigation

7 Requirements for Trustworthy AI	Associated Risk for the Insurance Sector (Risk Level)*	INDUSTRY PERSPECTIVE
R1 - Human agency and oversight	**Low**	**Risk mitigation:** - engaging with various stakeholders representative of different societal voices to reach consensus on fair and responsible use of AI applications, giving particular emphasis to business use cases with a potential impact on vulnerable customer (e.g. use of mandatory behavioral change schemes) - training and skill development programs to help internal staff to embed ethical considerations within evolving business practices
R2 - Technical robustness and safety	**Low to Medium**[34]	**Risk mitigation:** - implementation of state-of-the-art cybersecurity standards and control frameworks to ensure that AI applications are resilient against both overt attacks and more subtle attempts to manipulate data or algorithms
R3 - Privacy and data governance	**Low to Medium** (depending on the use case)	**Risk mitigation:** - disciplined internal oversight of the data science pipelines - strong business data ownership and accountability can support the accuracy of data. - High dependence between the level of governance of data and the model accuracy with the business purpose (e.g. for medical related services the scrutiny level must be much higher than for marketing purposes).
R4 - Transparency	**Low to Medium** (depending on the use case)	**Risk mitigation:** - developing and implementing internal guidelines and policies to ensure a consistent approach to the transparency and explainability of algorithmic outcomes. These guidelines need to assist the assessment (risks/benefits) on a case-by-case level. - When using more complex / non-linear AI models, to ensure wider customer acceptance it is critical to adopt highly interpretable and explainable techniques for non-technical audiences (e.g. optimal classification trees)
R5 - Diversity, non-discrimination and fairness	**Low to Medium**[35] (depending on the use case)	**Risk mitigation:** - enforcing internal guidelines and policies which continuously refine and validate use of AI applications against established ethical values, paying particular attention to sensitive customer characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious belief.
R6 - Societal and environmental wellbeing	**Low**	**Risk mitigation:** - designing AI applications that contribute to a positive and sustainable societal outcome, by ensuring that internal guidelines and policies consider the principles of solidarity and inclusiveness (e.g. social and financial inclusion, affordability) in the development of insurance products and services
R7 - Accountability	***Low***	**Risk mitigation:** - Robust internal governance with clarity on roles, responsibilities and accountabilities - Strengthened quality assurance for model oversight and controls so as to demonstrate that internal governance systems are robust and uncompromised enough to address challenges resulting from use of non-linear AI models.

34 Medium, if cybersecurity risks aren't properly mitigated.
35 Medium risk, for some use cases (e.g. Life & Health, hyper-personalization, pricing and underwriting) if fairness and discrimination aren't properly assessed at every step of the value of chain.

4.2 Recommendations for diverse Stakeholder segments

Recommendations to: developers/users of insurance AI

Some recommendations to insurers for the responsible and ethical use of AI within the organisations include the following:

- **Ethical AI corporate culture:**
  - o ***Internal guidelines and policies***: Insurers can mitigate AI risk by developing and implementing internal guidelines and policies that will ensure a consistent approach to the ethical and trustworthy AI, including the transparency and explainability of algorithmic outcomes.
  - o ***Ethical AI audits***: An internal ethical algorithm audit mechanism for the design, development and deployment of AI systems in insurance, could ensure that the moral and ethical issues surrounding the use of AI are being addressed, while identifying potential biases or flaws, depending on the type of industry and globally accepted auditing procedures and standards.

- **Ethical AI training:** The responsible design of AI entails a number of decisions that are made by engineers. As such it is important to raise awareness at different management levels through appropriate educational and training programs. Contributing this way to ethical decision making at an individual and corporate level and realising responsible AI-driven innovation-by-design.

Recommendations to: governments and regulators

- There is an urgent need to set up national and international regulatory frameworks to ensure democratic governance of artificial intelligence so as to prevent its misuse but at the same time facilitate responsible AI innovation.

- Particular attention should be given to potentially high-risk applications (e.g. robo-advice for asset allocation or long-term financial products), which could be considered for ex-ante regulatory approval.

- In order to ensure the adequate implementation of ethics into AI there is also a need to (re)examine the "emerging" role that the regulator of the future should have. The regulator of the future can be seen as an entity that works closely with business actors from the insurance and other sectors, in order to certify AI products and services while at the same time protect the public (WEF, 2019). In this perspective, the role of governance is shifting and expanding as a concept. The dynamics of AI bring about the need for an agile, anticipatory AI governance

approach. That is an adaptive, human-centred, inclusive and sustainable policymaking approach, anchored in the notion that policy development is no longer limited to governments but rather is an increasingly multi-stakeholder effort (WEF, 2018).

- Multi-stakeholder consensus should be reached on what constitutes a responsible use of AI and data. This involves the active engagement of the public, private sectors as well as research and academic stakeholders and the civil society at large, but also the *democratisation* of AI across these segments.

- At an infrastructure level, a big challenge for the insurance sector relates to the standardization of the exchange of data with 3rd parties such as banks (i.e. what kind of data is involved, etc.). Facilitating data exchanges is a crucial component of AI-driven innovation in insurance.

Recommendations to: customers

- Actively engage in discussions with diverse stakeholders and communities in order to express what constitutes a responsible use of AI and data from their perspective.

4.3 Generic Recommendations

An EU Ethical and Trustworthiness AI Label

Figure 2: AIEIG AI Ethics Label (Source: AIEIG, 2020)

The operationalisation of ethically-sound AI applications, could involve the creation of a European Union AI Ethical and Trustworthiness Index that could function as an "AI ethics label" (similar to the European Union Energy Label) that would rate and classify the different AI applications based on their AI Trustworthiness level (at a requirement level), while considering the contextual parameters. This rating could be implemented before the use of an AI system (and could even be used as a guideline during its design phase) and should be constantly monitored and updated. Furthermore, the required level to ensure that an AI application is ethical and trustworthy would be highly dependent on the application context (i.e. industrial context, B2C, B2B, B2P, etc.) and thus the associated risk of the current and future use of the system.

Such a labelling scheme would provide useful information, in

an easy to communicate way to the end users as they choose between products and services from different companies, to the European policy makers, regulators and standard-setting organisations to constantly update and enhance their requirements for ethical and trustworthy AI systems as well as to encourage companies to design, develop, implement and invest in ethical and trustworthy AI applications-by-design and -by-default.

An example of this approach is the VCIO model and AI ethics label that has been developed by the AI Ethics Impact Group[36] (AIEIG, 2020) which combines four conceptual parameters, namely values, criteria, indicators and observables in order to evaluate AI (Figure 2).

New models of Governance

Artificial Intelligence will profoundly change all industrial sectors, transforming our business and social interactions. Equally profound is and will continue to be our need for new principles, algorithms, and policies so as from the one side to accelerate the positive impacts of AI and from the other to minimise the negative (expected and unexpected) consequences. Towards this aim, new models of governance are needed that will move beyond traditional reactive governance towards anticipatory models of governance. Such innovative governance models will be agile and adaptive in nature. They will be human-centric, inclusive and sustainable by design (agile governance). Such models will be centered around collaborative, multi-stakeholder policy development processes, rather than government-centric functions. They will embrace change and empower rapid, on-going readiness and adaptiveness (proactively and anticipatory).

Balancing between top-down, bottom-up and middle-out AI approaches/policies

In order to seize the opportunities that Artificial Intelligence offers to society, we are confronted with a daunting challenge: how can we promote the development and deployment of AI-driven innovation and enhance existing business practices in the insurance sector as well as in other sectors, while at the same time limit the risks and failures? How can we ensure citizen protection-by-design, and increase trust and confidence in AI? How can we balance innovation and trustworthiness in AI, without compromising our ethical standards and fundamental rights? Furthermore, the lack of specific and verifiable AI principles and measures may threaten the effectiveness and enforceability of AI ethics guidelines.

36 The AI Ethics Impact Group is an interdisciplinary consortium led by VDE Association for Electrical, Electronic & Information Technologies and Bertelsmann Stiftung.

Analysis shows that a balanced approach could be of benefit for ensuring "ethically sound" design, development, implementation and evaluation of AI systems (by-design) in the insurance context among others. Aiming to find the framework that will provide the right mix that will from the one side prevent (proactively) the potential negative effects to fundamental rights and human-centric ethical standards and from the other side provide a future-proof and innovation-friendly framework. Consequently, such an approach would entail the effective integration of different types of legal regulation namely: '*top-down*' (*legal regulatory action*), '*bottom-up*' (*self-regulation*) *and* '*middle out*' (*co-regulation and coordination mechanisms for the governance of AI*[37]) policy actions, considering the contextual parameters of the different AI insurance applications and the associated risks (current and anticipated).
These different levels of legal regulation are aligned with AI4People's Report on Good AI Governance (Pagallo et al., 2019b), and include:

I. **Traditional legal regulation** - '*top-down*' *approach* - including both hard law and soft law actions. such as the Opinions of the Art. 29 Working Party and, nowadays, of the European Data Protection Board (EDPB) in the field of data protection, as a set of rules or instructions for the determination of every legal subject of a system. These are the rules that aim to directly govern social and individual behaviour, and mainly hinge on the threat of physical (and financial) sanctions as a means of social control (Kelsen 1949);
II. **Self-regulation** - '*bottom-up*' *approach* - including the different forms of self-regulation in all its variants. According to Directive 2010/13/EU, 'self-regulation constitutes a type of voluntary initiative which enables economic operators, social partners, non-governmental organisations or associations to adopt common guidelines amongst themselves and for themselves' (Recital 44)[38]. An example of such a bottom-up approach, in the context of AI, is the final Assessment List for Trustworthy AI[39] (ALTAI), presented by the European High-Level Expert Group on AI during July 2020, which is intended for self-evaluation purposes. Sector-specific considerations can add relevant elements to the ALTAI list, for each specific AI system. This type of actions acts in a complementary manner and does not substitute other legislative requirements.
III. **Co-regulation** - 'middle out' *approach* – incorporating both elements of top-down legal framing and bottom-up empowerment of individual actors. According to Directive 2010/13/EU, 'co-regulation gives in its minimal form a legal link

37 The 'middle-out' layer (Pagallo et al., 2019a; 2019b) includes everything that lies between the top-down and bottom-up approaches and is associated with forms of co-regulation.
38 Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive).
39 The ALTAI is also available in a web-based tool version, available at: link.

between self-regulation and the national legislator' (Recital 44)[40]. In this context, the regulatory role is shared between stakeholders and the government or the national regulatory authorities or bodies (Pagallo et al., 2019b). In the current EU legal framework, this "middle-out" layer is mostly associated with forms of co-regulation models, such as the GDPR (Pagallo et al., 2019b)[41].

A classification framework for AI applications

The following framework is based on Ziouvelou and McGroarty (2021). In order to facilitate the creation of novel AI insurance applications and services that will serve the common benefit and welfare and the customer needs in an ethical and trustworthy manner, a classification framework for AI systems and applications is needed.

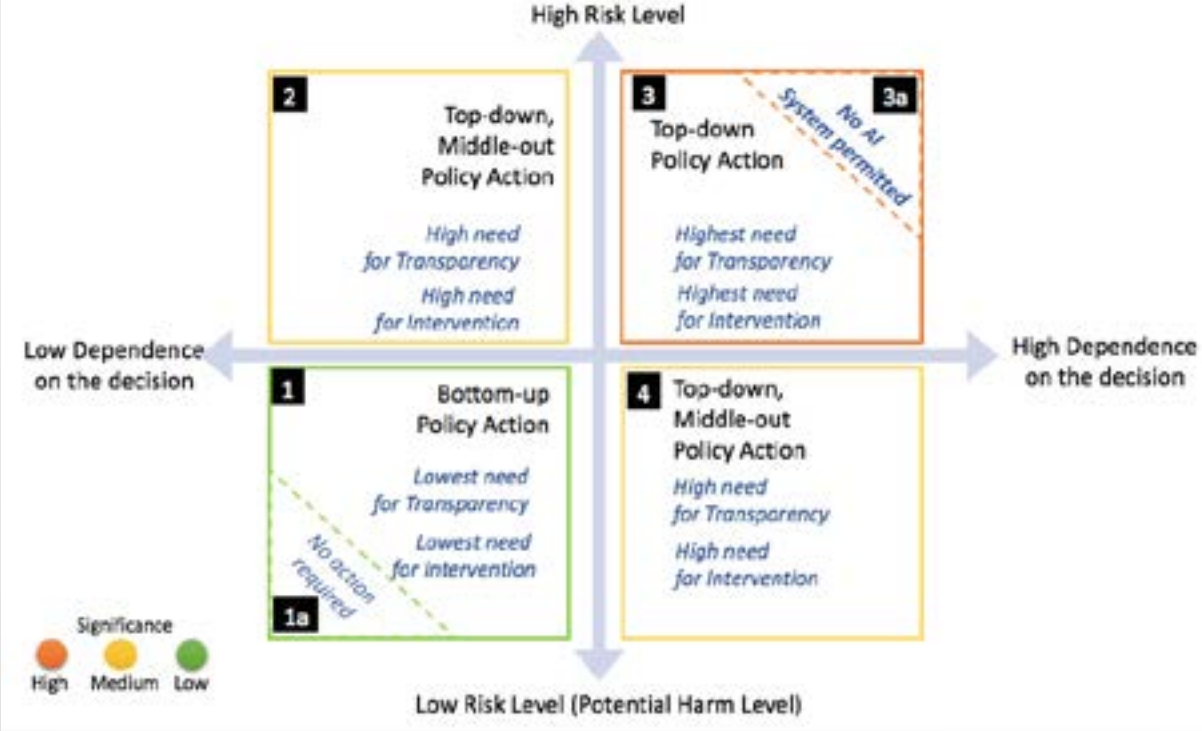
Based on the two axis that focus on the level of risk and the dependence on the decision or outcome of the AI application, four distinct categories emerge that have different levels of significance (high, medium, low), as shown in Figure 3. Furthermore, two additional subcategories emerge within Category 1 and Category 3 denoting the extreme and opposite cases, namely: *subcategory* 1*a* – where we have the lowest risk level and the lowest possible dependence on the decision and at the other end, *subcategory* 3*a* – where we have highest possible risk level associated with the highest possible dependence on the decision. The highest class, 4 in this case, serves to classify contexts where no AI system should be applied. For algorithmic decision-making systems that fall between these two extremes, a subdivision of at least three further classes seems to make sense to reflect increasing system requirements adequately.

- ***Category*** 1: AI systems that belong in this category have a low risk level and our dependence on their decision is low. This category entails AI systems of low significance that necessitate the lowest need for transparency and the lowest need for intervention. As such 'bottom-up' policy actions may sufficiently cover the trustworthiness requirements of applications that belong in this area.
- ***Subcategory*** 1***a***: This subcategory depicts one of the extremes as it entails AI systems that have the lowest possible risk level and the lowest possible dependence on their decisions. As such, these systems have a very low overall significance and an equally low associated impact for humans and nature. They can be parts or a basic stand-alone AI systems, which once assessed and classified in this subcategory they may necessitate no further action.

40 See previous note.
41 According to Pagallo et al., (2019b) in the GDPR "the mixed approach to governance revolves around the 'principle of accountability' that through a mix of primary and secondary legal rules aims to strike a balance between guaranteeing compliance with both the principles and the top-down rules of the system, while leaving room for self-regulatory measures."

Figure 3: A classification framework for AI applications/systems and associated policy action

(Source: Ziouvelou and McGroarty, 2021)



- ***Category*** 2: AI systems that belong in this category have a high-risk level and our dependence on their decision is low. This category includes AI systems of medium overall significance that necessitate a very high need for transparency as well as intervention. Consequently, they call for top-down legal regulations and co-regulation actions (middle-out).
- ***Category*** 3: This category depicts AI systems of high overall significance that exhibit a high-level of risk and our dependence on their decisions is high. Such systems require the highest level of transparency and intervention. Thus, top-down policy actions are needed for AI systems in this category.
- ***Subcategory*** 3***a***: This subcategory entails AI systems that exhibit the highest possible risk level and the highest possible dependence on their decisions. Consequently, these systems have a very high overall significance and impact for humans and nature. As such AI systems, assessed and classified in this subcategory they should not be permitted.
- ***Category*** 4: AI systems that belong in this category have a low-risk level and our dependence on their decision is high. Such systems have a medium overall significance and necessitate a high need for transparency and intervention. Consequently, they necessitate both top-down legal regulations as well as middle-out actions.

References

Spielkamp, M. (2019). Automating Society: Taking Stock of Automated Decision-Making in the EU. BertelsmannStiftung Studies, AW AlgorithmWatch gGmbH, 2019.

Centre for Data Ethics and Innovation (2019). AI and Personnel Insurance, CDEI Snapshot Series.

Council of Europe (2019). Unboxing Artificial Intelligence: 10 steps to protect Human Rights. By the Council of Europe, Commissioner for Human Rights, May 2019.

Deloitte (2015). Insurance Disrupted. General insurance in a connected world.

European Supervisory Authority for the Insurance Industry (EIOPA) (2019). Big Data Analytics in Motor and Health Insurance: A Thematic Review. Luxembourg: Publications Office of the European Union.

European Commission (2019). Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence (AI HLEG), April 2019. Retrieved from: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

European Commission (2020). On Artificial Intelligence – A European approach to excellence and trust. Retrieved from: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. High-Level Expert Group on Artificial Intelligence (AI HLEG), July 2020. Retrieved from: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Fang, Lei, Gianvito Lanzolla, and Andreas Tsanakas. "Digital Technology Adoption and Changes in Management Priorities." Academy of Management Proceedings. Vol. 2020. No. 1. Briarcliff Manor, NY 10510: Academy of Management, 2020

Financial Conduct Authority - FCA (2018) Pricing practices in the retail general insurance sector.

Financial Conduct Authority - FCA (2018), "Pricing practices in the retail general insurance sector: Household insurance", Thematic Review, TR18/4, October 2018.

Financial Conduct Authority - FCA (2019), General insurance pricing practices, Interim Report, Market Study, MS18/1.2, October 2019.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707.

International Association of Insurance Supervisors (IASIS) (2016), Issues Paper on Cyber Risk to the Insurance Sector, Financial Crime Task Force, August 2016.

Kellogg, Katherine C., Melissa A. Valentine, and Angele Christin. "Algorithms at work: The new contested terrain of control." Academy of Management Annals 14.1 (2020): 366-410.

Lanzolla, Gianvito, Danilo Pesce, and Christopher L. Tucci. "The digital transformation of search and recombination in the innovation function: Tensions and an integrative framework." Journal of Product Innovation Management (2020).

Larsson, S., Anneroth, M., Felländer, A., Felländer-Tsai, L., Heintz, F., and Cedering Ångström, R. (2019). Sustainable AI: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence. AI Sustainability Center.

OECD (2019). Recommendations of the Council on Artificial Intelligence (adopted on May 22, 2019). Retrieved from: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

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

Pagallo, U., Aurucci, P., Casanovas, P., Chatila, R., Chazerand, P., Dignum, V., ... and Valcke, P. (2019a). AI4People-On Good AI Governance: 14 Priority Actions, a SMART Model of Governance, and a Regulatory Toolbox.

Pagallo, U., Casanovas, P., and Madelin, R. (2019b). The middle-out approach: assessing models of legal governance in data protection, artificial intelligence, and the Web of Data. The Theory and Practice of Legislation, 7(1), 1-25.

TOGAF, (2017). An introduction to the European Interoperability Reference Architecture (EIRA) v2.1.0. Retrieved from: https://joinup.ec.europa.eu/sites/default/files/distribution/access_url/2018-02/b1859b84-3e86-4e00-a5c4-d87913cdcc6f/EIRA_v2_1_0_Overview.pdf

The AI Ethics Impact Group (AIEIG), (2020). From Principles to Practice, An interdisciplinary framework to operationalise AI ethics. Retrieved from: https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf

The Bribery Act (2010), UK Ministry of Justice. Retrieved from: http://www.justice.gov.uk/downloads/legislation/bribery-act-2010-guidance.pdf

The Geneva Association (GA) (2020). Promoting Responsible Artificial Intelligence in Insurance. Jan 2020, International Association for the Study of Insurance Economics.

The Guardian (2016). Facebook forces Admiral to pull plan to price car insurance based on posts, by Graham Ruddick. Nov. 2, 2016. Retrieved from: https://www.theguardian.com/money/2016/nov/02/facebook-admiral-car-insurance-privacy-data

Universal Declaration of Human Rights. Retrieved from: https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf

World Economic Forum (WEF) (2018). Agile Governance: Reimagining Policy-making in the Fourth Industrial Revolution. WEF White Paper, April 2018.

World Economic Forum (WEF) (2019). A Governance: A holistic approach to implement ethics into AI. WEF White Paper Jan. 2019.

Ziouvelou, X. and McGroarty, F. (2021), A classification framework for AI applications/systems and associated policy action. Working paper, Centre for Digital Finance, University of Southampton.

6
LEGAL SERVICES INDUSTRY

Authors

Burkhard Schafer
Innovation Officer and Research Scientist, Institute of Informatics and Telecommunications, National National Centre for Scientific Research Demokritos, & Member of the Scientific Committee on Data Policy and Artificial Intelligence, National Council for Research and Innovation (NCRI), Greece

Cornelia Kutterer
Senior Director, Rule of Law & Responsible Tech, European Government Affairs at Microsoft

Elisabeth Staudegger
Professor at Universität Graz, Austria

Evdoxia Nerantzi
Policy Manager, European Government Affairs at Microsoft

Jacob Slosser
Carlsberg Foundation Postdoctoral Fellow at University of Copenhagen, Denmark

Jamie J. Baker
Associate Dean and Director of the Law Library; Professor of Law at Texas Tech University School of Law, USA

Mireille Hildebrandt
Research Professor on 'Interfacing Law and Technology' at Vrije Universiteit Brussel, Belgium

Rónán Kennedy
Lecturer in Law, School of Law, National University of Ireland Galway, Irelands

1. Introduction: scope and remit of the report

1.1. Methodological and terminological preliminaries

The aim of this report is to distil a number of ethical principles to assist developers of 'legal technology' and the users (such as public administration decision-makers and the legal profession) who commission it. We are addressing a specific AI-driven technology that is intended for use in the legal sector, giving 'AI' a broad interpretation that includes symbolic AI and rule-based systems as well as legal ontologies, legal information retrieval or data driven approaches such as machine learning and neural networks. Since the terms 'legal tech' and 'legal technologies' are now commonly used, we will use these to refer to those uses of AI that are germane to the delivery of legal services and the justice system, excluding 'generic' technologies that merely assist lawyers with tasks, such as billing and other back office activities. Our focus is on the use of artificial intelligence in the sense of symbolic logic developed to translate legal norms into computer code, as well as machine learning developed for legal search or e-discovery. We use 'legal technology' with quotation marks to emphasise that we mean the use in the legal domain, rather than an evaluative attribute of the digital artefact (legal vs illegal or ethical and unethical AI).

One challenge for such a project is the sheer scope of applications of AI in the justice system. They range from intelligent information retrieval tools to decision support systems in sentencing; from automated compliance for business processes ('code as law') to AI tools that help visualising evidence in trials; from chatbots that help citizens to file a complaint or generate documents to tools that predict the outcome of cases for litigation risk assessment. Many of these tools can be framed under the heading of 'legal search', and are already integrated in legal practice. Without claiming a comprehensive overview we direct the reader to examples such as Westlaw Edge, Lexis-Nexis and many more that are specifically relevant in Europe. While some of the recommendations below will be more relevant to some such applications than others, for all of them responsible use requires a clear focus on the moral underpinnings of law and the rule of law. This is why our focus is on the justice system in the broad sense of that term (including legislature, judiciary, public administration, public prosecutor, police and professional legal advice), because through them and their interdependencies law and the rule of law acquire concrete form.

On the technological side, the wide range of applications means that a wide variety of approaches to AI are used, from 'good old-fashioned AI' (GOFAI) in chatbots, to machine learning in prediction-based approaches, each with their own risk profiles and ethical challenges. On the application side, this means that what is at stake for individuals can differ dramatically, from determining a prison sentence to missed

opportunities to file a successful compensation claim for a minor product defect, or an automated check if all data used in a business process have appropriate consent forms associated with them. While the severity of the impact is one element to take into account in any assessment of the impact on law and the rule of law, two things are important to remember. First, in the justice system, there is no fixed and context independent matrix of 'severity'. To stay with the example, getting legal aid for small claims litigation may be a major issue for those on a lower income, and have a similar impact on them as a high impact commercial litigation could have on a well-resourced party. Harm, in both cases, may go beyond the economically tangible and deprive them of the dignity they deserve as citizens and human beings. Second, ethical assessment cannot stop at the evaluation of single transactions or use-events. Some of the most problematic ethical consequences of technology in the law will accrue when such missed opportunities accumulate, or small disadvantages become a systemic and endemic feature.

The wide scope of AI applications in the legal domain also means that the central questions of any ethics of AI, i.e. 'how are existing power imbalances, injustices and biases perpetuated, amplified or challenged, and what new power relationships will emerge due to new players entering the field?', may have quite different answers in different contexts. Transactional law, for example, a lawyer using a software tool for the drafting of contracts for well-resourced parties with access to independent high-quality legal advice, obviously poses very different ethical issues from a program that determines the eligibility for legal aid or social security payment for indigent claimants, or which supports decision making in criminal trials.

Technology is *never* neutral. Every use of AI tools in the legal field can have consequences that impact law and the rule of law. Legal information retrieval systems may lead to a loss in access to the law for some groups, or – as often claimed – to an overall increase of access to justice. It can marginalise uncomfortable but necessary minority opinions, or amplify them. Computational decision support in criminal trials can amplify or reduce human biases in decision making. Document assembly systems can reduce costs for indigent litigants, but can also encourage frivolous litigation that increases demand for legal solutions and hence costs for the courts, or be used as a pretext for the government to reduce legal aid even further. One key principle that we propose below is a 'rule-of-law risk assessment' that is sensitive to these various parameters.

As noted above, this report focuses on those issues that set law as a domain apart from other applications. It should therefore be read in conjunction with the general guidelines on responsible development and employment of AI that AI4People has developed. These lay out principles that any ethically sound development of AI ought

to observe, while this document will focus on duties that are specific to the legal domain and create obligations that go above and beyond those ethical concerns identified in the first two AI4People reports that apply to all uses of AI.

It is also important to reiterate that ethical considerations start when legal compliance has been achieved. None of the ethical principles stated in what follows should be misunderstood as an alternative to compliance with applicable laws, or as pre-empting further regulation. Aspiring to ethical excellence is only possible when it is driven by genuine concern for the wellbeing of others, not when it is instrumentalised to prevent appropriate control through democratically legitimised regulators backed up by sanctions. This is particularly important for applications that may have severe implications for the fundamental rights of citizens. Where such an infringement can be foreseen, appropriate legislation and institutional safeguards rather than calling for adherence to ethical guidelines will be necessary.

Finally, we need to remain mindful of the danger of technological determinism and solutionism. As we will see below, some of the current debates surrounding ethical use of AI have a long history, a history that also traces failed attempts to solve difficult issues of fairness through merely improved technology. We should therefore not start from the assumption that the adoption of a proposed 'legal technology' is a given and the only concern is to shape them in ethically less harmful ways. Rather, the question of their introduction always requires keen attention to the preliminary questions (1) what problem they supposedly solve, taking into account (2) what problems they will not solve and (3) what problems they may create. Deployment should not be the default. Spiekermann-Hoff talks in this context of 'the art of omitting', the prudential decision to not develop now, or not to develop at all, a certain application of a technology.

1.2. AI in Law as a unique challenge for AI ethics

The use of artificial intelligence in the justice system creates significant opportunities to address known shortcomings and failings in the administration of justice, but also poses unique and serious dangers, not only for individual citizens, but for the rule of law ideal more generally.

Any attempt to build an ethically responsible approach to the development, deployment and post-deployment monitoring of AI in the justice system and the legal services industry thus faces a number of challenges that are specific to the legal domain. An ethos of 'move fast and break things' is ill at ease with conceptions of justice and the rule of law that evolved slowly over centuries, and which are anchored in complex social understandings of law, justice and its role in democratic society.

At the same time, recent events have also brought into stark focus that not all aspects of legal traditions are worth preserving, and that many of them perpetuate deep injustices.

Technology can play a positive role in addressing these problems. This is not limited to legal institutions that do not work as intended or became corrupted, but also concerns legal institutions that work as intended and are imbued with particular ethical significance. Sometimes normative justifications of existing practices have been back-projected on legal institutions that arose for very different historical reasons; sometimes practices and institutions arose as the result of ethical concerns that were valid at the time of their creation, but are now obsolete. This means that not every conflict between 'legal technology' and the internal logic of legal institutions, or their functioning, is *necessarily* wrong. Any claim that a practice or institution serves justice deserves scrutiny and criticism. Even such a hallowed and central proposition such as 'equality before the law' is known to lead to substantive injustice when applied too formalistically – the famous prohibition preventing rich and poor alike from sleeping under bridges, or a prohibition on either men and women breastfeeding in public. At the same time, legal institutions in democratic societies do reflect deeply held ethical commitments and communal practices centred around shared values. An appropriate ethical assessment of AI in law therefore needs both – a willingness to critique established features and institutions of the law that are in need of change, and respect for the values and ethical commitments that are inscribed in many of these practices and institutions.

Legal systems across the world struggle with high costs of litigation and efficient enforcement of rights, and with often unconscionable delays in the administration of justice so that justice delayed does indeed often become justice denied. Even in rich countries, large parts of the population are often excluded from affordable legal advice. Absent a realistic ability to challenge discriminatory or otherwise wrongful application of the law, abusive practices against vulnerable citizens can become institutionalised.

Here technology can be a force for good, and not using its full potential in itself becomes an ethical failure. Law regularly deals with citizens who are in particularly vulnerable conditions, and subject to severe power and information imbalances. The employment of AI in law could reduce these imbalances by making information more easily accessible, but could also amplify and entrench them further into the very infrastructure of the administration of justice, making them even less responsive to social needs and more difficult to change through the political or judicial process.

Legal systems are also shaped by the need to reconcile deeply held and conflicting intuitions about justice. The application of general legal norms articulated in natural language inevitably requires interpretation and more or less discretion, which is however itself guided by the underlying principles of the rule of law. The dual imperative

of 'equality before the law' and 'doing justice to particulars' (Bankowski and others), following the rules but where appropriate showing mercy and empathy, thus requires a further balancing act that ensures the adaptiveness and the contestability of the law. The antinomian character of this dual imperative is a feature rather than a bug (Radbruch), and we must be mindful of the danger that the success of even a beneficial and, on its own terms, ethically sound legal technology risks 'resolving' this antinomian character by silencing or marginalising competing intuitions about the just society and stifle more substantial social change. Finally, while other fields of AI application can look at law as an external means to control abuse, the transformative effects on law itself make this more difficult, and raise the stakes substantially. Were AI systems to undermine the law, its ability to control their application would be lost.

Public excitement about new technologies risks that lessons from the past can easily be forgotten. Legal systems by contrast are always also, for better or worse, repositories of past experiences and a collective memory of problem solving. While there has been a recent upsurge in interest in legal AI, the idea has a long pedigree, with early systems developed in the 1970s and 80s. While most of them were expert systems that focussed on explicit, symbolic representation of and reasoning with rules, applications of machine learning can also be found from the 1990s onwards, for instance in the Split-Up system developed by Zeleznikow. In 1996, the UK became the first country to legislate for fully automated decision making in the Social Security Act. Revisiting some of the ethical debates of the time is instructive, both to see the continuity of ethical concerns that were raised then and now, but also to be reminded of issues that are not as visible in the current debate.

Looking at the parliamentary debate, we can identify three different types of concerns:

a. How can the legislature or the executive be certain that the computer program faithfully replicates the law – are the decisions it reaches correct?
b. If they are not, how can citizens know, and what are the remedies at their disposal?
c. Even if the decisions were correct in all cases, and could be communicated adequately to the citizen, isn't the decision to use it for social benefit applicants, and only for these, sending a symbolic message that some of the most vulnerable members of society do not count, that the state has given up on them and now only manages them through a cold machinery, where it should embrace and integrate them fully?

It is especially the third aspect of legal technology that is often missing in contemporary debates, which focusses on questions of technical correctness, debiasing of algorithms and explainable AI. It reminds us of the fact that simple technological solutions alone will not resolve complex social problems. While improvements in our ability to 'get the results right' are of course welcome and important, and also create an ethical obligation to use the best available tools, there remains a very real danger that they will nonetheless amplify and entrench existing biases and injustices. The principles therefore try to reach a more holistic appraisal of 'legal technology', and urge caution of their use in situations where the value of 'being judged by one's peers' also means 'being judged by one's fellow humans'.

Despite great expectations in the transformative potential of 'legal technology' in the 80s and 90s of the previous century, and significant uptake in some applications such as tax law, the first wave of 'legal AI' did not result in the substantial transformation of the justice sector that some had anticipated (see e.g. Susskind). Many of the current technologies share many of the features of these earlier attempts. Legal chatbots represent knowledge in very similar ways to earlier expert systems, data-driven pattern analysis in court decisions, and earlier ideas of modelling discretionary legal decision making. Undoubtedly, there have been significant improvements in the technical tool kit available to developers (see Bench-Capon 2015 for a historical timeline), both in terms of software and algorithms, and the hardware necessary to process large data sets at ever increasing speed. However, some of the reasons why this time we may see a much greater impact on the justice system, have less to do with changes in the way legal knowledge is represented and made computational, and more with wider social and technological developments, some of them with ethical salience. More data is 'born digital', including data generated by the justice system. This creates new opportunities but also dangers for access to justice, as commercial AI systems can also create walled data gardens. Coding has become easier, and in particular, platforms that support 'no-code automation' enable many more people to build simple legal applications. While the development of earlier systems typically took place in university environments or particularly large and well-resourced law firms or public administrations, this has now become possible for small start-ups, individual practitioners or citizen-activists. Platforms such as Josef promise 'legal automation for everyone' and have significant potential to support the work of NGOs and other activists that try to address shortcomings in the current justice system. While this democratisation of 'legal technology' has significant potential to contribute to the 'good AI society', it also means the loss of gatekeepers, bringing with it problems of quality control and rogue players. The principles account for this through a risk assessment framework that matches the danger of harm with a range of duties of quality assurance. Finally, the ubiquity of smartphones together with developments such as online banking has increased both our willingness and expectation to carry out more complex tasks which would previously

have involved expert assistance on our own, at a time of our choosing. It has changed the landscape of the discourse on digital exclusion and the digital divide, but not eliminated it. This means great care has to be taken to ensure that the underlying infrastructure for legal technology does not amplify inequalities and exclusion, something at least as important as the 'fairness' of the algorithms themselves, and inseparably connected with it.

2. Foundational Principles for Responsible use of AI in law

In this section we develop a series of principles that concretise the specific ethical concerns that should inform both the development and the use of AI in law on a high level of abstraction that applies to the whole range of technologies and use cases. As they concern the foundations of law and the rule of law they cannot be reduced to ethics, indeed they concern the nature of modern positive law as informed by the rule of law.

2.1. Principle of Respect for the integrity of law and the rule of law

We identified respect for the rule of law as the key ethical concern that sets law apart from other domains of AI application. It creates a 'way of thinking' or internal logic that is central to legal reasoning and is not always aligned with the way AI developers conceptualise the world. With 'integrity of the legal system' we mean respect for law as a separate discourse that cannot be replaced or taken over by other modes of thought. We must therefore first clarify and motivate what we mean with the centrality of this concept, and its dynamic and often contested nature.

Different technologies 'fit' in different ways to competing conceptions of justice and competing legal philosophies. Widespread adoption of comparatively simple rule-based expert systems and similar technologies that first entered the scene in the 1980s and 90s, and face a resurgence through legal chatbots on the one hand, 'compliance support tools' on the other, may seem to enhance formal equality, but could shift the balance towards formalism simply because they are unable to accommodate more nuanced and discretionary reasoning. This can lead to injustices in cases that do not fit their pre-defined categories. Data-driven approaches, by contrast, could in theory allow decision makers such as judges or public sector administrators to do 'justice to particulars' and consider much more nuanced and much more varied factors in their decision making than currently possible, and for instance tailor sanctions much more to the individual circumstances of a case. However, 'personalised law', when taken to the extreme, conflicts with the promise of equality before the law, may destroy social bonds and allows extraneous factors to influence outcomes.

Every legal system tries to balance these conflicting visions, equality and predictability, through rule adherence and responsiveness to individual differences and contexts. One overarching danger of legal technology, even when its outcomes are seemingly benign, is that substantive social debate and democratic deliberation about the nature of justice and how to achieve it are pre-empted by what is technologically possible. This can be seen in proposals of 'mechanising' legal and administrative decision making, not because there is a desire in society for a more uniform application of rules that passed public debate and scrutiny, but because this is the only thing that is currently technologically feasible, or worse, feasible given cost constraints.

Furthermore, once computing infrastructures are put in place at significant cost, these may be resistant to change in line with social attitudes, again side-lining the democratic decision-making process. This also refers to the difference between (1) legality as part of the moral backbone of the rule of law, which is intimately linked with the argumentative nature of the law that nevertheless provides for legal certainty (Waldron), and (2) the computational legalism that emerges in the context of both rule-based and data-driven AI in law (Diver).

For the purpose of this report, this poses one of the multiple challenges that need to be taken into consideration when assessing the ethical implications of a proposed technology. Preserving the integrity of a legal system (Dworkin), its internal logic and values, is an important consideration. At the same time, disrupting structures that perpetuate injustice can be ethically mandated, even though it can raise in turn questions of democratic legitimacy and accountability if done by software developers or their clients.

Turning these abstract ideas into more concrete actionable rules, we have the following overarching **Principle of Respect for the integrity of law and the rule of law:**

> *Any use of AI must respect the integrity of the legal system, the values inscribed therein, and adhere to practical and effective respect for the rule of law. Disruptions of the internal logic of law are permissible only if they in turn are justified by an overriding ethical mandate.*

2.2. Principle of purposiveness

As both use and non-use of AI in the justice system may create risks for individuals and society, anticipatory evaluation of potential impacts are central for ethically robust development of AI in law. From this we can derive an overarching general requirement of anticipatory impact assessment: Any decision on the development and deployment

of AI in law needs to start with an assessment of potential risks for the values of the justice system, for the groups likely to be adversely affected, and, where applicable, for the human rights of those who will interact with the technology, be it as subjects of a decision, as users, as data engines or as legal knowledge experts. As a requirement of the democratic state under the rule of law, such impact assessments should be publicly accessible to allow scrutiny by civil society.

We do not adhere to the innovation mantra that the burden of proof is on those opposing the introduction of 'legal technology'. We therefore propose that the first stage of any impact assessment consists of a clear and comprehensible answer to the three questions related in the introduction. We name this the principle of purposiveness, which should be integrated in the anticipatory impact assessment:

Principle of purposiveness:

Any proposed legal technology should be explicitly upfront about

(1) *what problem it supposedly solves,*
(2) *what problems it will not solve*
(3) *what problems it may create.*

Only with this degree of transparency at the initial stage of a project can a meaningful ethical evaluation take place. This evaluation then has to determine, in addition to the utility of the technology to achieve these goals, the following:

a. How will the technology in the short term impact on the character of all the stakeholders that are affected; how does their role and self-understanding change?
b. Which human, social, economical etc values are positively or negatively affected?
c. What are the essential societal values and priorities that need to be protected from this development?

2.3. Principle of respect for (the situated nature) of the rule of law

This principle inevitably poses the question of the standard against which such an impact assessment should take place. Even within Europe, for historical reasons, there is substantial divergence regarding how legal systems operate, and how justice is best served.

To give an example, in some legal systems, the jury is seen as essential not just for efficient decision making, but as a requirement of justice. A random selection of citizens

who can decide without fear of repercussions, and (therefore) without publishing reasons, is seen as the only way to constrain the exercise of arbitrary power by the executive, creating public acceptability of and trust in the justice system, and ensuring fairness to the individual by giving them at least the chance to be judged by people who as their peers can relate to their context, values and life decisions. For legal systems without a tradition of jury trials, or systems with recent memory of abuse of the 'popular voice' in the administration of justice, juries can appear as the opposite. The randomness of their selection, the 'black boxed' nature of their reasoning, and the fear of manipulation of emotional responses or bias creates fears very similar to those levelled against (some forms of) AI.

Whether or not a proposed future AI application for law raises ethical concerns and conforms with Principle 1, it will also depend on these conceptions and understandings of justice. An AI that profiles jurors to determine what arguments they are most likely to be susceptible to would be a serious ethical concern in those systems that imbue the anonymity and secrecy of the juror deliberation with normative qualities, but neutral (irrelevant) or even beneficially increasing transparency in those systems that followed a different trajectory, inscribe different historical experiences and consider black box character of jury deliberations an ethical concern.

Conversely, systems that profile judges and their preferences may be seen as particularly problematic in those continental systems that consider legal formalism also as an ethical mandate, as they potentially introduce extra-legal considerations into the way in which lawyers plead their cases. This may explain the recent French ban on such systems – but may be considered ethically beneficial from the perspective of those legal systems that consider jurisprudential realism not just as a descriptive account of the operation of law, but as an aspect of sound and competent legal advice that lawyers owe to their clients. Similarly, even within a system that values formal rational decision making as embodiment of justice, profiling judges can be beneficial if it were to show unjustifiable biases in the administration of justice – not to exploit them for the benefit of an individual client, but to criticise and remedy them.

There are thus no universal precepts for AI ethics in legal technologies. Instead, the principles of democracy and the rule of law require jurisdiction specific, context sensitive concretisation. However, the European Human Rights framework, in particular the right to a fair trial, the right of judicial review of administrative decisions, and the rights to privacy, freedom of information and non-discrimination provide non-negotiable boundaries for the margin of discretion that national jurisdictions have to concretise legal protection. This allows us to formulate a principle of respect for the situated nature of rule of law:

Principle of respect for the situated nature of the rule of law:

The use of AI in law ought to take account of historically grown, socially and culturally embedded practices of adjudication, and divergent conceptions of justice within the contours of the European fundamental rights framework.

Legal harmonisation can have of course benefits. But the decision to align legal practices across borders has to be driven by democratically legitimated decisions, not by consideration of convenience of transnationally operating 'legal technology' companies promoting a single product across diverse legal markets and traditions.

2.4. Principles of fair distribution of impact at individual and societal level

Legal AI does not divide neatly into ethically sound applications that help to address known shortcomings on the one hand, and harmful applications that create unjustifiable risks on the other. Often, the long term social and ethical consequences of 'legal technology' will be difficult to anticipate, as legal professionals, governments and citizens in turn adjust to the new technology.

To illustrate by way of example, at first sight, developing a tool that gives reliable, fast and free legal advice to underserved groups, seems like a clear case of beneficial use as described in the first paragraph. However, the availability of such tools could in turn be used to justify a further reduction in legal aid, which may affect also those cases that due to their complexity are less well suited for this technology and potentially threaten a hard- won achievement of the rule of law, the right to counsel. In the long run, this could lead to an overall reduced access to qualitative legal support for indigent claimants, for whom even inferior software products are deemed 'good enough'.

Conversely, faster and cheaper access to tools that facilitate litigation can on the one hand give people the ability for redress that they had been excluded from previously. It can however also lead to a more litigious society, and far from reducing the burden on the legal system, create an increase in demand.

At least one clear ethical duty applies to both developers and those who commission such tools. Where the development of a 'legal technology' is driven mainly or exclusively by a cost reduction rationale, there is a particularly strong duty to communicate transparently, explicitly and honestly any trade-offs in terms of quality, accuracy, comprehensiveness and similar quality criteria that apply to professional legal advice.

Especially where efficiency threatens to trump quality, an impact assessment should mark out who will benefit from cost reduction and who will pay the price in

terms of legal protection. Such an assessment should include an assessment of potential mitigation strategies that can move risks away from already vulnerable or disadvantaged groups.

Taking risks cannot mean taking risks with other people's rights and interests, particularly in the legal services industry.

This can be further concretised by the

Principle of fair distribution of risks and benefits:

'Legal technology' is shaped in contexts with significant power differentials. Responsible development and employment of AI reflects on these structural conditions, and protects against unfair redistributions of risks and benefits. More ambitiously, AI in law should aim to reduce existing power imbalances and redistribute risk to those best placed to mitigate it.

The application of this principle will be highly context dependent. However, if linked explicitly to the human rights framework, we can ensure that the protection of the 'awkward minorities', or on computational terms, the 'difficult edge cases' are always foregrounded.

The rationale of the human rights framework is to protect individuals and minority groups from the 'dictatorship of the majority' and from exploitative business models of private enterprise alike. Justice is a public good that cannot be traded as if it were part of the logic of economic markets. For AI tools that are based on statistical pattern recognition and analysis, this is a particular challenge, due to the implications of mathematical optimisation that may obscure outliers and result in the tyranny of 'stochastic majorities'. Responsible AI in law will require a more fine-grained analysis to ensure equal respect and concern for each individual citizen (Dworkin) at every step of the design process.

Where historically disadvantaged groups or atypical individuals are put at risk, the introduction of a 'legal technology' should be reconsidered if effective safeguards and remedies cannot be provided. Community involvement is crucially important here, enabling those who will suffer the consequences of 'legal technology' to voice their concern and contribute their lived experiences – designing *with* rather than merely for them.

This leads us to the problem that sometimes, an AI tool can be ethically neutral, and even beneficial, when used by an individual lawyer for the benefit of an individual client, and yet cause significant harm once the use of the very same system becomes

widespread or universal. In these cases, the AI fails to abide by the Kantian imperative (and also Jonas' extension of this principle to techno-social systems) that demands that unless an action can be universalised, it is not ethically sound.

To illustrate this point, consider a system that on the basis of past court decisions predicts the likely outcome of a court case. A lawyer using such a system may use it to discourage a client from bringing litigation, or contest a case, where there is insufficient chance to prevail. This can protect the client from unnecessary expenses or worse. It can also relieve the pressure on the justice system that can then use scarce resources for more meritorious cases, also benefiting others as a result. Despite these apparent beneficial traits, general use of the same algorithm could lead to highly undesirable consequences for the justice system and the rule of law. Some fact constellations would become unlikely to reach the courts, simply because in the past, cases like these were unsuccessful. As a result, the legal system ossifies and becomes irresponsive to social change.

This means we need to extend the **principle of fair distribution of risks and benefits with a principle of transversal impact assessment**, that anticipates risks for the common good of practical and effective legal protection:

> *The rule of law and the concept of legality transcend the binary relation between individual and state, or lawyer and client, and constitutes a common good. Evaluating ethical risks of employing AI in law must therefore consider long term detrimental impact on third parties and the cumulative effect on our ability to live lawfully.*

3. Principles for Responsible Development of AI in Law

Legal technology does not just raise ethical issues at the point of application, they also raise issues in the way they are developed. Unlike the laws of nature (including such things as rules on how to diagnose an illness, one of the earliest expert systems), legal rules are not true or false, but authoritative or not authoritative. Their authority, especially in democratic societies, in turn is closely linked to the way in which they are generated in a rule-governed legislative process. Ethical issues can arise when program developers shortcut or usurp this legislative process.

In particular in applications that serve as 'compliance tools' and either strictly enforce or at least nudge individuals and businesses into law-conforming behaviour, ethical issues arise from the moment when software developers translate legal rules into code. 'Regulation through software architecture' (Lessig) as a new form of governance through legal technology raises its own set of ethical concerns that we will turn to now.

3.1. Principle of procedural transparency

As the expression 'Code is Law' indicates, we can think of some forms of 'legal technology' as a new form of regulation. However, the process of translating natural language into computer code inevitably also involves, like every translation, a process of interpretation and change of meaning. Ambiguous legal phrases may have to be disambiguated, decisions have to be made on how much of the original meaning needs to be operationalised and how much must be left analysed in the parameters, and choices will have to be made if implicit logical connections in the text ought to be made explicit in the code. The problem here is not so much that the code may be wrong – this issue will be dealt with elsewhere – but that the programmers have to take on tasks that in the past were reserved for the legislator (e.g. clarifying laws through statutory instruments or codes of practice) or the courts (interpreting and disambiguating legislation).

This raises questions about the legitimacy of such law-making by private actors, the danger it poses to the process of democratic control of legislative and executive powers, and the possibility for citizens to participate in a deliberative democracy. We therefore argue that legislation by way of code is inherently creating ethical risks and can be incompatible with the core tenets of democracy and the rule of law. First, it can violate the procedural precepts of good law-making that ensure public debate and accountability. Second, it can create new access barriers to the law. Special skills are now needed to understand what the law says, and also to determine whether a given application is a true representation of the law. The use of arcane and complex language in natural language law making has been recognised by initiatives such as the plain English campaign. The values of open administration of justice and equal access to the law, which drove this development towards more accessible legal language, risk being undone through code-based compliance tools, if these do not also afford interrogation in natural language.

Even though every form of automated decision-making is in this sense carrying high risks, we propose a set of principles that should apply whenever delegated legal powers are exercised by way of policy rules whose application may be automated.

H.L.A. Hart, in *The Concept of Law*, distinguishes between primary rules of conduct and secondary rules that regulate how primary rules are created, enforced and given legitimacy. 'Legal technology', especially in the form of automated decision and compliance systems, tends to incorporate primary rules only, which threatens to cut the nexus between laws and the conditions that ensure their legitimacy. Responsible development of 'legal technology', by contrast, is aware of the wider constitutional and social rules that confer legitimacy to laws, respects the values enshrined in constitutional

settlements, and abides by the procedural rules and safeguards that control ordinary law-making.

We focus here in particular on two of Hart's rules, the Rule of Recognition and the Rule(s) of change. Hart describes this rule as 'to say that a given rule is valid is to recognize it as passing all the tests provided by the rule of recognition and so as a rule of the system. We can simply say that the statement that a particular rule is valid means that it satisfies all the criteria provided by the rule of recognition.'

In democratic societies, this means in particular that the law emanates ultimately from the legislature and elected officials, albeit in some cases indirectly though powers that have in turn be conferred according to constitutional law. It can also mean that procedures about public involvement were observed, and that the law maker can be held responsible both through democratic processes and through judicial review by the courts. AI developers are not subject to democratic and court control, nor are they recognised as legitimate norm-givers. Where AI is used to automate the application of legal rules, there is therefore a heightened demand for transparency and accountability, to ensure that developers do not by accident usurp functions that only 'recognized authorities can exercise'. Equally, the right of the public to participate in legislation and hold legislators to account must not be diminished, and legislators in turn must not be able to avoid public discourse and scrutiny by turning political questions into technological design decisions.

Just as the traditional legislative process creates an auditable paper trail from committee discussions to plenary debates to the eventual drafting of the law that can be used in judicial review, so should legislation through coding. This is the rationale behind the

Principle of procedural transparency:

Throughout the development process of 'legal technology', decisions about design features that are functionally equivalent to interpretation, augmentation or limitation of the law ought to be documented, including a documentation of who made the decision and on what authority. This should happen in language accessible to all stakeholders, including civil society and their representatives, and not just specialists.

3.2. Principle of respect for the legislative process

Connected to this notion of transparency is the next principle that we propose as a key aspect of a new theory of prudential digital law-making by code. It aims to mirror the procedural safeguards, limitations on the exercise of power, and the right of civil

society to contribute to legislative processes that we find in traditional law making.

This may mean at some point in the future the creation of a new *sui generis* system for legislative drafting, including the possibility by legislators to 'legislate in code', and similar formal rules that would change the way in which modern states generate laws. In the absence of such rules, a number of ethical constraints can however be identified that flow from the Rule of Recognition and the principles of democratic accountability and the sovereignty of the demos.

Principle of respect for the legislative process:

Public sector organisations that commission legal technology must not use it as a way to prevent democratic rule-making, scrutiny and accountability, or limit established rights of the public to participate and be heard in the legislative process. They remain ultimately responsible that 'legal technology' matches in form and function the laws it implements.

This aims to prevent the often-observed tendency by public bodies to change the nature of a problem from one that requires public debate to one that is turned into a technical coding issue resolved behind closed doors by specialists. It also aims to avoid acting *ultra vires* by delegating legislative power to commercial software developers, and to ensure transparency and public scrutiny of the norm-making process.

3.3. Principle of unfettered public participation

The purpose of the previous principle was to protect public participation rights in the legislative process. The underlying value is expressed directly in the

Principle of unfettered public participation:

The right of the public to contribute in the norm-setting process, and the right of communities to be heard in public sector rule-making that affect them, must not be reduced or circumvented by a process of building legal technology and 'legal by design' environments that would rule out disobedience.

Participatory, community-led design processes will frequently be superior to achieve responsible 'legal technology' that aligns with human rights and democratic principles. In most public sector 'legal technology' projects, it is an ethical demand. But it will frequently be desirable also in most other AI initiatives in the legal domain. Ethics for the responsible development of 'legal technology' cannot be reduced to a set of rules, in the way a software specification manual might. Rather, responsible 'legal technology' is fundamentally a set of processes that ensures the voices of all affected individuals and groups are heard and attentively listened to.

3.3.1. Principle of transparent and adequate compensation

Participatory design processes can be resource-intensive. They can also place significant burdens on the communities and groups it tries to involve. This is exacerbated by the fact that the most vulnerable communities are regularly also facing economic and other barriers to efficient political engagement and contribution. An ethical emphasis on community-led design practices and other forms of substantive involvement of affected parties should not further increase the burden on them. To be not a mere aspiration, but a reality, also requires the provision of adequate resources, compensation, and recognition for the labour of these groups. One key achievement for modern democracies was the introduction of payment for parliamentarians. This provided the infrastructure that was needed so that active and passive votes also aligned in practice. From this historical experience we get the

Principle of transparent and adequate compensation:

Development and employment of AI in law does not just passively listen to the voices of all individuals and groups that are affected by the operation of 'legal technology', it actively seeks and involves these voices. Through their labour these communities add value to the eventual product, value and labour that has to be adequately compensated and acknowledged.

3.3.2. Principle of diversity and representation

In some situations, substantial community involvement will not be possible or necessary, for instance in very low risk applications. Even then, there should be independent ethical scrutiny and advice where possible. While having diverse and representative ethical oversight is a requirement for all AI development, it takes on particular importance for AI applications in the domain of law. Principles of representativeness, non-discrimination, inclusivity and substantial as well as procedural fairness are also core requirements of the rule of law in democratic societies. This is expressed in the

Principle of diversity and representativeness:

The groups and individuals who through their labour and expertise inform the development of 'legal technology', also and in particular through ethical advisory boards and other governance structures, should be representative of the society that the technology will serve and reflect its diversity.

3.4. Principle of non-discrimination

While the previous principles focused on the process of creating 'legal technology', the question of diversity, discrimination and bias takes on particular importance in legal contexts. That we do not deal with this issue more prominently is due, first, to its importance for *all* AI applications and not just the domain of law, and it has therefore been addressed already in the previous AI4People reports. Second, the previous principles that require transparent decision making, and principles that require that diverse voices are heard in the design process, should already minimise the danger of products that could lead to biased or discriminatory decision making by its users. However, due to its centrality for the rule of law, and also because there can be different forms of discrimination, it is restated here explicitly. It is connected with the rule of recognition in that biased and discriminatory legal practices threaten the very foundations of a legal system.

Principle of non-discrimination:

> *Biased and discriminatory practices are incompatible with the rule of law ideal and its promise of justice for all. 'Legal technology' must demonstrably ensure through the choice of proper design methodologies (e.g. vetting of input data), choice of technology (algorithms that are interpretable and/or have been debiased), testing (both during design and after deployment), and an analysis of other forms of disparate impact on communities, informed by their lived experience that they do not unjustly discriminate against individuals and communities.*

This principle goes beyond mere mitigation against biased decision making, such as reliance on data that encodes social prejudices in, for example, the detection of crime or the imposition of penal sanctions. We include other forms of potentially discriminatory practices, such as for instance exclusion from digital legal information due to aspects of the technology that make it unusable for people with disabilities. Equally covered by the principle can be graphic design choices, for instance for the avatar of a legal chatbot, which currently more often than not reflect stereotypical assumptions about gender roles ('assistants' tend to have voices, images and animations that present female, while 'legal expert AIs' such as ROSS present male voices and appearance etc), physical attractiveness or are chosen predominantly from major ethnic groups. This technological amplification of existing social biases is undesirable, even when used in the private sector, but it becomes an ethical afront when used by public sector 'legal technology'. As a default, anthropomorphising 'legal technology' should be avoided or minimised to facilitate the exercise of critical judgement by the citizen who is interacting with them.

Ultimately the principle of substantive equality before the law flows from the overarching human rights principle of human dignity. Of all the human rights ideas this is the most difficult to concretize, and as we saw in the historical introduction one of the most persistent concerns in 'legal technology' that is driven by a cost rationale.

3.5. Principle of temporal contestability

Finally, we turn to a second meta-rule identified by Hart, the Rule(s) of Change. Rules of change ensure that legal systems do not remain static, but react to change. While there are different approaches across Europe, especially when it comes to constitutional law, they all contain limits on how much, and for how long, a current legislator can bind their successors. Rules of change describe the necessary process and the powers that are conferred to various actors to affect such a change.

The danger with 'legal technology' as part of the infrastructure is that it can be resistant to change, creating unwieldy legacy systems that are difficult or impossible to change, modify or abandon even when the democratic process or judicial review by courts make this necessary. The most prominent example of this problem in recent times has been blockchain technology and the degree of immutability of rules inscribed on the chain that they bring.

This gives us the

Principle of temporal contestability:

All 'legal technology' projects should anticipate that laws change in response to changing societal norms. 'Legal technology' must not lock in or constrain future legislators, and must permit challenges and changes to current laws through the democratic process or through judicial review.

'Designing for change' means avoiding design decisions that impact on parliamentary sovereignty and remaining adaptive and flexible at least to the same degree as traditional laws are.

4. Principles for Responsible Employment of AI in Law

The next set of principles looks at the point in the lifecycle of 'legal technology' where the technology is rolled out and in use. Now, the problem of potentially wrong advice, decisions or recommendations becomes relevant. The first group of principles looks at the use of 'legal technology' in public administration, the second focuses on uses by the court system. The principle of continuous empirical evaluation also applies to legal practitioners, and for them will be taken up and concretised in the Principle of Collective Responsibility.

Although some principles are more relevant to particular fields of application, and are presented in this thematic fashion, they are nonetheless valid for all practical deployments of AI in law. The public sector is responsible for its use of AI, and must ensure that this use remains justiciable. Practising lawyers are ethically accountable for the tools that they adopt, and must ensure that they support access to justice. The judiciary must remember that they have a role in monitoring their own responsibility and accountability for the adoption of AI in courts, and ensuring that the public and private sectors are aware of ethical expectations in this regard. The boundaries between the categories in what follows are porous, and the ethical duties listed apply to all involved in the application of 'legal technology', but some may be more pressing in specific contexts.

4.1. Accountability: Principles for Responsible Legal AI in Public Administration

The previous principles addressed ethical concerns with 'legal technology' that would have to be considered even if, hypothetically, the technology works perfectly. The next set of principles deals by contrast with one of the most obvious concerns, the possibility that a given program comes to incorrect results and as a consequence causes harm, including but not limited to the denial of a right to the subject of a decision.

4.1.1. Principle of continuous empirical validation

One of the characteristics that sets law apart from other applications of AI is that what counts as a 'correct' answer can be radically contested in legal settings. How do we know that the system performs correctly, at least on its own terms? If a citizen has been harmed through an incorrect decision that was caused by the use of technology, how would they be able to know? Assuming that no system is perfect, when are the benefits so great that they make taking the risk of occasional mistakes ethically acceptable? How fine grained should that analysis be? Can it be acceptable to use a system that in most cases outperforms a human decision maker, but systematically

makes more mistakes when dealing with the claims of a small subgroup? We will revisit these issues below, but we note as a general principle the

Principle of continuous empirical validation:

Ethical use of AI in law requires rigorous and realistic testing both pre- and post-deployment, with robust methods to share findings about problems between users, and users and regulatory agencies.

The last part of this principle, the mandate to share findings, will also play a role in the Principle of Collective Responsibility that we will discuss below. We do note here though that this principle and its corollaries have been successfully used in medical research and development for a long time, without imposing impossible burdens on the industry once appropriate systems are in place. In medical research, pre-registration of tests, and meta-analysis of shared results through e.g. the Cochrane review, are well established, as are post-deployment systems to report adverse effects of drugs. Similar mechanisms to share experience are needed at least for those data-driven AI applications in the justice system that carry higher risks. 'Higher risk' applications include all those where AI systems directly assist decision makers such as judges.

While the demands made of AI systems in law regarding their accuracy will vary between contexts (we can and should judge legal information retrieval systems differently from sentencing support systems, and both from a chatbot that guides citizens through a process of form filling), benchmarks need to be created that allow those who use and commission legal technology, and those adversely affected by it, to evaluate the reliability of 'legal technologies' in terms of real world reliability (rather than accuracy on the data). For those applications that imply risks for fundamental rights and freedoms, this benchmarking ought to be carried out by independent bodies.

4.1.2. Principle of contestability

As noted above, in law, the question of what counts as the 'correct answer' can be in turn contested. This means that it will not be sufficient to certify that the AI has some desirable formal properties. Rather, decisions must be explained in a way that allows their contestation. What type of explanation or justification do we owe to the citizens who are affected by the use of an AI system? What other communicative duties are there – for instance towards a possible supervisory or appeals body, or especially in cases of trials and similar procedures to the public at large?

In contemporary debate, these issues are often conceptualised as an aspect of the correctness of the decision-making, and are seen as a technical challenge that can be

addressed through interpretable or explainable AI. If the subject of an automated decision, or a user of legal technology, can get an explanation of its decision, mistakes can be identified and, in theory, addressed and rectified.

In the legal field, this idea is closely connected to the adversarial nature of many legal settings, especially when they are also embedded into a hierarchical system of judicial review and appeals. Legal decision making, in many contexts, centres around a dual notion of 'contestability'. Horizontal contestability means that in an adversarial setting, both sides deserve to be heard, to bring their arguments to the table, examine and challenge each other's position. Vertical contestability means that legal decisions can regularly be challenged and submitted to a higher authority. This can take the form of judicial review in administrative decisions, or appeals in trial situations. AI in law, especially in the case of 'black-boxed' machine learning systems, can potentially undermine this central conception. To counter this, we postulate the principle of contestability as a concretisation of the principles to respect the internal integrity of the legal system and the rule of law ideal.

Principle of contestability:

The use of AI must not limit the right to contestation of a decision where the legal system provides such a right. Contestation requires typically that appropriate reasons for a decision are given.

4.1.3. Principles of justification and explanation

'Explainable AI in law' is one aspect that supports the right to contestation. 'Explainable AI' typically makes the way in which a decision was reached transparent. However, in the legal field we are not just interested in how a decision was reached. Rather, we want a *justification* that meets certain standards and criteria. In epistemology, this is sometimes expressed as the difference between the context of discovery and the context of justification. We do neither expect nor desire that human judges explain a decision in terms of their personal background, socialisation or life history, as important as these may be for the way a decision was reached. Neither do we demand an explanation in terms of neurological data from the judge's brain at the point of decision making. Rather, we demand a justification that meets certain legal standards, e.g. by reference to the applicable law and governing cases. From this we get the

Principle of adequate justification:

For applications of AI in law to be ethically defensible it is not sufficient that 'an' explanation of the decision is given. Rather, to be valid the explanation has to respect the

internal logic of the legal system in question, and give appropriate reasons that match the duties to justify decisions that we currently impose on human decision makers such as judges or public administrators.

However, we may allow in some cases also to contest a decision that can be justified in this way, but where the procedure to reach the decision was irredeemably flawed. A decision by a judge who failed to recuse themselves even though they were related to one of the parties would be a typical example. In cases like this, 'back engineering' a formally valid justification of the outcome of the decision is not enough and cannot remedy the flaw in the decision-making process. Similarly, we should expect from AIs also to be explainable in the sense that flawed reasoning processes can be identified, and where appropriate remedied, even in cases where the outcome (by chance) is correct.

Principle of sufficient explanation:

AI in law must be explainable enough to allow adequate allocation of responsibility to enable sanctions and redress where there are errors or misconduct and to vindicate the duty to provide remedies in those cases where AI applications may have caused unjustified harm. Where, exceptionally, such an explanation is not possible, comprehensive no-fault compensation or similar safeguards are necessary.

The aim of this principle is preventing any 'responsibility gap' and to ensure that parties harmed by an AI are not facing burdens of proof they cannot discharge.

4.1.4. Principle of practical and effective redress

Neither explanations nor justifications in the above sense are sufficient on their own to address concerns about false decisions. Knowing that the decision of an AI was wrong, and why it was wrong, do not in itself right the wrong that was committed. There is a danger that 'transparency' is weaponised to shift responsibility from the software developer to customers, users or the people subjected to autonomous decision making. Especially in legal contexts and the differences in power, material and intellectual resources and social capital that often permeate them, knowing that an injustice was committed is often not sufficient if there are no practical and effective means to obtain redress. This leads to the

Principle of practical and effective redress:

Any evaluation of the ethical soundness of a 'legal technology' project needs to identify the remedies that will be provided to challenge decisions, to ensure that incorrect decisions

can be detected, making sure that these remedies are adequate to provide redress and serve as a deterrent to negligence or misbehaviour, and that those harmed are given appropriate information and support to exercise their rights effectively.

4.2. Justiciability: Principles for Responsible AI in the Judiciary

A particularly sensitive use of AI in the justice sector is its use by courts and the judiciary. This is not just due to the high stakes for the citizens involved, but the social and symbolic value of 'seeing justice being done'. The next principles therefore add a number of principles that are specific to AI in the courtroom.

4.2.1. Principle of public justice

It would be a mistake to reduce the issue of explainable AI to the detection and correction of mistakes. Even if, hypothetically, it could be formally proven that a given AI application in the domain of law was always 100% correct, there would still be a need to give reasons for its decision in many contexts.

The trial in particular is a process of 'public holding to account' (Duff and Duff). In giving public reasons for a decision, a judge, or a similar decision maker in a public body, does not just communicate the reasons for the decision to the party directly affected, to enable them to contest the decisions. The aim of the trial is also to enable them to understand, and in some way accept, why justice demanded this outcome.

Furthermore, by giving *public* reasons, trials make the control of the justice system a collective duty and possibility for all citizens, particularly by making it easier to detect instances of corruption, bias, and prejudice. Trials also publicly reaffirm the values of the community, and in this way also allow criticism and evolution of the law in response to changing social attitudes. 'Government with the consent of the governed' requires involving the public in both, scrutiny of the application of legal rules in a given case, and scrutiny of the acceptability of the legal rules given the results that their application leads to. The idea of the public trial is thus not only central to the rule of law, it also intertwines democratic participation and citizenship.

'Legal technology' can undermine this principle in a number of ways. It can weaken the obligation of state actors to explain and justify their decisions, for instance when they use software owned and controlled by a private sector enterprise that shields behind trade secret or copyright law. The above principle was meant to address this. It can also exclude citizens from their role as observers of the justice system, by using communication tools that are inaccessible for them. AI can, in these situations, act as an assistive technology, that supports citizens with their participation, through

automated translation for instance, or as assistive technology for citizens with visual or hearing impairments. Conversely, it can limit access to and publicity of the trial – for instance when for the purpose of evidence evaluation, judges or jurors 'see' a crime scene re-enactment in their virtual reality helmets, an experience that cannot be replicated for observers (Cieslak).

Finally, it can create economic obstacles of accessing legal information through business models that rely on intellectual property rights over these legal documents. From this we derive the

Principle of Public Justice:

Public justice, notably the publicness of the trial, is a cornerstone of the fair trial. The use of 'legal technology' is ethically harmful if it increases the obstacles for the public to get involved with, and learn about, the justice system. Responsible legal technology, by contrast, aims to lower these obstacles, while preserving legitimate privacy interests of those involved in the proceedings.

4.2.2. Principle of equal access to justice

As noted above, one important aspect of the idea of public justice is access to legal materials. One reason why Lon Fuller's 'King Rex' in his seminal 'The Morality of Law' failed to build a legal system was by preventing access of citizens to the legal texts that ruled them. In our context, these can be primary sources generated by the public sector such as statutes and regulations, court decisions, policy documents, or documents generated by parties such as submissions, pleadings and briefs. The idea that as a very minimum, citizens need to be able to determine what the law is, i.e. have access to primary legal sources, remains patchy, unaffordable for many and subject to numerous obstacles in practice. Technology has significant potential to mitigate this problem. Many legal systems have made substantial improvements of publishing statutes and also, in varying degrees, court decisions online. Intelligent and user-centric Information Retrieval systems can help even those with less experience to find material relevant to them. Automated translation tools can give access to the law to recent immigrants and speakers of minority languages.

Technology can, however, also create new barriers, and increase inequalities for access to justice. These can be caused by design features of the technologies, such as accessibility problems caused by inadequate consideration of the needs of users with a wide range of disabilities, digital skills, and unequal access to computers. They can also be created through a combination of business models and IP law, when walled gardens

are created for legal information. The above publicity principle can therefore be sharpened to

Principle of equal access to the law:

'Legal technology' should strive to maximise access to legal sources for all. It should in particular be used to widen access to groups that have historically faced significant access barriers. It must not lead to new obstacles, including technological and economical obstacles. Design for accessibility that considers a wide range of disabling factors, in close consultation with the affected communities, is central for ethical 'legal technology'. Alternative modes of access have to be preserved for those who cannot be accommodated this way.

Because access to the law is mediated by language, this principle also speaks to the principle of respect for local traditions that we introduced earlier. Law and language are intimately connected, as are languages and the way in which cultures express themselves. In the European context, this has been recognised in the European Charter for Regional or Minority Languages, Article 22 of the Charter of Fundamental Rights that prohibits discrimination on grounds of language, and institutions such as the European Language Equality Network. Just as there is a danger that 'legal technology' could impose conceptions and solutions that work only for some (typically larger) legal systems on others, there is a concern that speakers of majority languages are better served in fields such as natural language recognition. At the very least users of recognised minority languages must not be disadvantaged by the introduction of 'legal technology'.

The ability to speak one's own language when interacting with the state, the legal system and public authorities, and in turn get services in one's own language, is an important aspect of full democratic participation for speakers of recognised minority and regional languages. For our purposes, it connects the question of 'legal technology' more broadly to that of participation in the democratic process, the relation between citizen and legislature and the ideal of full and active citizenship.

4.3. Responsibility: Principles of Responsible AI in Legal Practice

Apart from their use in the public sector, the legal profession is the main user of 'legal technology'. The unique characteristics of the legal profession constitute another factor that sets AI in law apart from many other domains. This needs to be accounted for in a suitable ethics framework.

In most jurisdictions, practicing lawyers are members of a regulated profession, which necessarily also means they are subject to the ethical and professional rules of

their respective regulatory body, which is instituted by law. These rules in turn can be enforced with sanctions, and must be qualified as legal norms (delegated in the relevant Act of Parliament). Some of the principles in this document could become part of these rules and in that way obtain legal status. This highlights the special nature of professional ethics in the domain of law as 'ethics rules with teeth' that blur the line between ethical rules and enforced regulatory standards. As the regulatory bodies and professional societies will have to play a critical role in ensuring beneficial use of AI, some of the principles directly address the regulatory bodies, as they will have to play a central role in the development, use and monitoring of AI applications in law.

4.3.1. Principle of equivalent application of professional ethics rules

This principle reflects not just the centrality of law and the rule of law for democratic societies, and with that the particularly severe risks that can be created. It also reflects the special function that lawyers play in society as a regulated profession, and the dual private/public status they have in many jurisdictions as both representatives of their parties but also officers of the court. This function has always been understood as generating a set of standards of 'professional ethics' and concomitant obligations that go above and beyond those of ordinary citizens or unregulated businesses. As an example, lawyers have a duty of confidentiality *vis-à-vis* their clients that goes above and beyond those owed by all businesses under data protection law. A particular ethical concern with the use of AI in the legal domain is that while law is a regulated profession, computing, in the main, is not, or at least not currently. This creates dangers for the users of 'legal technology' when these are not operated by regulated lawyers, but are created and provided by commercial entities that currently do not provide similar levels of protection and redress. From this we get the

Principle of equivalent application of professional ethics rules:

Where 'legal technology' takes on roles traditionally carried out by lawyers as members of a regulated profession, the protection for the affected citizen should not allow any circumvention of the core principles that govern professional conduct of practicing lawyers, notably (but not limited to) the norms set out by the International Bar Association concerning: independence, honesty, integrity and fairness, conflicts of interest, confidentiality and professional secrecy, clients' interests, lawyers' undertaking, clients' freedom, property of clients and third parties, competence and fees.

This will also include, for instance, availability of indemnity insurance and access to independent complaints mechanisms of the type often provided by the law societies and similar professional organisations. Any danger to misrepresent the interaction with a 'legal technology' as privileged legal advice, such as the use of legal symbolism, an

avatar with the insignia of a lawyer etc, must be strictly avoided. This principle should not be misunderstood as buttressing the monopoly of lawyers for legal service delivery. It does however ask for equivalence in protection for people who receive legal advice through a machine, whoever the owner-operator of the program may be.

4.3.2. Principle of ultimate responsibility

Regulated professions are distinguished by the autonomy that their members enjoy: 'professionals are autonomous insofar as they can make independent judgments about their work' (Bayles). This means, in particular, the freedom to exercise their professional judgement. However, this autonomy is not unlimited or self-serving. Professional autonomy can only be exercised in an ethically sound way if members of the profession 'subject their activities and decisions to a critical evaluation by other members of the profession.' (Hoogland and Henk). From this understanding of law as a regulated profession, we can derive two further principles.

Principle of ultimate responsibility:

The choice to use or not to use a given AI as a tool is an exercise of professional judgement. As members of a profession, lawyers have particular ethical responsibility for the tools they choose for the discharge of their responsibilities.

This includes a duty to carry out sufficient training, and remain up to date, to understand the potential, limitations and risks associated with the AI tools they use. Where lawyers rely on third party certification, they have a duty to understand the limitations of these parties, including the independence of the certification body.

4.3.3. Principle of collective responsibility

The previous principle does not entail that developers of AI software are freed from their ethical duties. Rather, both their obligations and that of the lawyers using an AI exist in parallel. Nor does it mean that lawyers need to acquire an understanding of AI that equals that of the experts who build the system. It does mean however that the choice for or against an AI tool has ethical and professional salience and creates a responsibility that cannot simply be shifted to the developers or suppliers. In practical terms, it means lawyers using AI need to have appropriate and current knowledge, training and understanding of the system that they are using, demonstrated e.g. through continuous professional development activities. For law societies and similar regulatory bodies, this means that competency in AI has to become an element of professional competency requirements. Part of this competency has also to be an understanding of one's limitations in knowledge, when it is safe to use technology despite these limitations,

and how and where to get the right type of support. Law societies and other such bodies whose independence from software manufacturers is guaranteed by statute are also well placed to develop systems of certification that can help their members in deciding on appropriate and safe tools that advance instead of diminish law and the rule of law.

The central role of professional bodies is furthermore reflected in the

Principle of collective responsibility:

The safe and ethically sound use of AI in the legal domain is a collective responsibility of lawyers and their professional bodies. This means a duty to share experience especially of problems and errors encountered.

This principle flows from the requirement (noted above) that members of a profession should 'subject their activities and decisions to a critical evaluation by other members of the profession'. Post-market reporting of problems with drugs in the medical field could be a model for this type of sharing, which could be administered by the professional bodies. These are also best placed to determine the right balance between the transparency and openness of this process, and the necessary confidentiality that enables frank disclosure.

4.3.4. Principle of technological neutrality

Being member of a regulated profession does not only create additional professional duties towards one's clients. The European Union Directive on Recognition of Professional Qualifications (2005/36/EC) defines professions as 'those practiced on the basis of relevant professional qualifications in a personal, responsible and professionally independent capacity by those providing intellectual and conceptual services in the interest of the client and the public'. Crucial here is the last part that establishes a duty to the public at large. Lawyers as officers of the court, and professional organisations as state-recognised regulators, have duties that go beyond the relation between them and their clients. They have responsibility towards the integrity of the legal system as a whole. Ethical problems can arise in particular where there is potential conflict between the ethical and professional duties owed to a client, and those owed to the general public.

From this we derive the

Principle of technological neutrality:

If lawyers or government officials employ 'legal technology' to provide a legal service or to make legally relevant decisions, their duties as professional lawyers remain unqualified, and do not shift to the system developers. The use of technologies does not alter what those subject to law may expect of legal professionals.

This principle does not say that technology is ethically neutral. In our view, 'legal technology' never is. Rather, it mirrors the principle of technological neutrality in Internet law, that states that what is illegal offline remains illegal online. Together with the Principle of respect for the democratic process and the Principle of equivalent application of professional ethics rules introduced above, it ensures that the use of AI in law maintains consistently at least the same standards and protections that we expect from human decision makers. In addition, it acknowledges that 'legal technology' may require more and different legal protection compared to that prevailing in the offline world. This implies that technological neutrality may require compensatory measures to ensure equivalent protection.

5. Concluding remarks

Most of the AI for People's 'global AI frameworks' have been structured on the seven key requirements for trustworthy AI, proposed by the High-Level Expert Group on Artificial Intelligence (HLEGAI): (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability. The framework for the development and use of AI in law has not followed this lead in its structure, to account for the diversity of the legal domain. Nonetheless, each of the Principles ultimately speak to this primary framework. They emphasise the professional responsibility of lawyers and their regulatory body (HLEAGAI 1), propose a system of evaluation and post-deployment review (HLEAGAI 2), emphasise the specific legal demands on explainability (HLEAGAI 4), foreground the role of law in promoting a vision of equal citizenship (HLEAGAI 5), discuss the need for a robust system of 'making good' in case of harm (HLEAGAI 7). Together, they protect the Rule of Law as cornerstone of a democratic society, and thus ultimately all serve HLEAGAI 7. Because of this central connection between a functioning justice system and the democratic ideal, the principles do not just address a narrowly defined 'legal services industry' but rather target legal practice in all its forms (law firms, the judiciary and the office of the public prosecutor, the legislature and public administration), software developers and digital publishers in the legal domain, and also those subject to a jurisdiction and thus entitled to legal protection against the violation of their human rights.

The reason for this resistance to limit the ethical principles for legal technology to either developers or commercial legal practice should be clear: the introduction of AI into the practice of law cannot be left to the legal services industry alone, and the underlying assumptions that legal protection is a commodifiable service whose price

should depend on negotiations in a competitive economic market is a category mistake. Fair, transparent and effective economic markets themselves depend on the legal systems that institute, sustain and develop their constitutive constraints (such as contract and property, but also including public law obligations and the institution of an independent judiciary capable of deciding and enforcing such constraints). Law therefore defines the contours and inner workings of economic markets and should not be framed as a commodity itself, dependent on the whims of a market meant to decide its fate.

This, it should be made clear, applies to both the decision to use or develop a suggested technology, and the decision *not* to use or develop a suggested legal technology. Justice systems across the globe are at breaking point. Years of underfunding after the financial crisis of 2008, and a raised awareness of systematic and historical injustices, have resulted in an erosion of public trust. Appropriate technologies have significant potential to increase access to justice, reduce economic, educational, geographic and cultural barriers to access legal advice. They can help monitoring and evaluating the performance of key stakeholders, enable new forms of public participation in the justice system, and democratise access to legal sources as an enabler of civic participation. Both of the ethical obligations to use/develop and not to use/develop a 'legal technology', as appropriate for the particular context in which a lawyer finds themselves, require reflection on the underlying values that the rule of law embodies, which is what this report ais to facilitate.

Moreover, whereas the law is informed by a specific set of ethical principles, notably justice, legal certainty and purposiveness, it differs from ethics in not depending on the ethical inclinations of whoever has the power to decide (the sovereign). The law is firmly grounded in the ability to unilaterally impose its will upon those subject to its jurisdiction, which makes the idea of depending on the ethical intent of those who serve the law rather hazardous. The paradox of the rule of law means that legal protection against the state in the end depends on the police force of that same state, and that this can be effective and meaningful if checks and balances have been instituted. Without countervailing powers in the heart of constitutional democracies, there is no rule of law but rule by law by persons (the reign of unconstrained sovereignty). A global AI framework should commit to upholding the rule of law and to preventing the development and use of AI for a more effective rule by law by persons (usually called a dictatorial regime).

Countervailing powers and checks and balances, however, require careful crafting. They demand a proper understanding of the integrity of the law, which entails both its consistency and its adaptiveness as integrated in the concept of legal certainty. This necessitates keen attention to legal norms as providing foreseeability, accessibility and proper safeguards that protect against rigid or unreasonable application. Law is an

argumentative practice that not only decides conflicts based on binding, publicly available rules and principles but also enables an adversarial debate over the correct interpretation of such legal norms. The advent of AI may cater to dreams of a self-driving law that is both unambiguous and personalised. The task of the lawyer is to resist such fundamental misunderstandings about the nature of law and the limits of technology. The lawyer's task is to work towards an integration of AI that supports checks and balances and maintains the rule of law rather than making efficiency the holy grail of legal services.

That being said the guidelines offered above do interact with the key requirements summed up below. **The integrity of the law** that is typified by the principles of purposiveness, respect for the rule of law and a fair distribution of impact, imply acuity with regard to the principle of technical robustness (if a 'legal technology' does not operate as claimed it cannot not serve the law's purpose), transparency and accountability (core to respect for the rule of law), and the principle of diversity, non-discrimination and fairness (which stipulates a fair distribution of impact). As to **development of AI in law**, we have asserted the principles of procedural transparency, respect for the legislative process, unfettered public participation, transparent and adequate compensation, diversity and representation, non-discrimination and temporal contestability. These connect with human agency and oversight (which assumes procedural transparency to be meaningful, as well as respect for the legislative process, unfettered public participation, diversity and representation to be truly inclusive as to human agency, and temporal contestability to ensure effective oversight), with technical robustness (which requires procedural transparency to ensure mathematical verification and empirical validation as well as diversity and representation to prevent training on data sets containing distributions that are the result of unfair bias), with privacy and data governance (as they require procedural transparency to detect infringements of privacy; respect for the legislative process to prevent government from being replaced by governance and non-discrimination and temporal contestability to assert that not just any data governance regime is fit for purpose in a constitutional democracy), with transparency (both in a procedural sense and in the substantive sense of enabling adequate compensation), and with diversity, non-discrimination and fairness (as explicitly addressed under diversity and representation as well as non-discrimination). As constitutional democracies need to cope with environmental and societal well-being, developers of 'legal technologies' should ensure that their effectiveness does not depend on disproportional impact on the environment (which aligns with the principle of respect for the legislative process that should not be hijacked by data- or code-driven approaches that endanger the sustainable development goals (SDGs) as advocated by the UN). Obviously the principle of accountability only makes sense in the context of a jurisdiction capable of holding those who violate its legal norms to account, whether big players, government or individual citizens. As to the **employment of AI in law** we

have detected eleven principles that are constitutive of legal practice in a constitutional democracy, grouped under the headings of accountability (focusing on the use of responsible AI in public administration), justiciability (focusing on the use of responsible AI with proper judicial oversight) and responsibility (focusing on the use of responsible AI in legal practice). We refer the reader to the content of the relevant chapter for a detailed analysis of the relevant principles, such as those of justification and explanation, practical and effective redress, equal access to justice and equivalent application of the rules of professional ethics. Many of these principles can be correlated one way or another with the HLEGAI global AI guidelines, but again, they are not merely ethical principles but mostly binding legal precepts and their granularity provides for concrete legal protection rather than ethical guidance.

Mapping the 'key performance indicators' of the HLEGAI against our own framework demonstrates many cross-references (and many more can be detected). This does not imply, however, that we could have done equally well by framing the use of AI in law based on those principles. We would have started on the wrong footing by assuming that general principles of the development, deployment and use of AI are fit for purpose regarding the architecture of constitutional democracies. Instead, we have started by laying the foundations (integrity of law), subsequently addressing the relevant challenges for developers, followed by stipulations for the actual employment of AI in legal practice. This should pinpoint the road forward, refusing to accept 'legal technology' by default (technological solutionism), while taking a pragmatic and principled approach to the integration of 'legal technology' in legal practice (opting for responsible AI).

Summary of Principles

Foundational Principles for Responsible AI in law

1. **Respect for the integrity of law and the rule of law**
   Any use of AI must respect the integrity of the legal system, the values inscribed therein, and adhere to practical and effective respect for the rule of law.
2. **Principle of purposiveness**
   Any assessment of the ethical valence of a proposed legal technology should be explicitly upfront about
   (1) *what problem it supposedly solves,*
   (2) *what problems it will not solve*
   (3) *what problems it may create.*

3. **Principle of respect for (the situated nature) of the rule of law**
   The use of AI in law ought to take account of historically grown, socially and culturally embedded practices of adjudication, and divergent conceptions of justice within the contours of the European fundamental rights framework.
4. **Principles of fair distribution of impact at individual and societal level**
   'Legal technology' is shaped in contexts with significant power differentials. Responsible development and use of AI takes account of these structural conditions, and protects against unfair redistributions of risks and benefits. More ambitiously, the use of AI in law should aim to reduce existing power imbalances and redistribute risk to those best placed to mitigate it.
5. **Principle of transversal impact assessment**
   The rule of law and the concept of legality transcend the binary relation between individual and state, or lawyer and client, and constitutes a common good. Evaluating ethical risks of using AI in law must therefore consider long term detrimental impact on third parties and the cumulative effect on our ability to live lawfully.

Principles for Responsible Development of AI in Law

6. **Principle of procedural transparency**
   Throughout the development cycle of 'legal technology', design decisions that are functionally equivalent to interpretation, augmentation or limitation of the law ought to be documented, including a documentation of who made the decision and on what authority. This should happen in language accessible to all stakeholders, including civil society and their representatives, and not just specialists.
7. **Principle of respect for the democratic process**
   Public sector organisations that commission the development of legal technology must not use it as a way to prevent democratic scrutiny and accountability, or limit established rights of the public to participate and be heard in the legislative process. They remain

ultimately responsible that 'legal technology' matches in form and function the laws it implements.

8. **Principles of unfettered public participation**
   The right of the public to contribute in the norm-setting process, and the right of communities to be heard in public sector rule-making that affect them, must not be reduced or circumvented by a process of building legal technology and 'legal by design' environments that would rule out disobedience.
9. **Principle of transparent and adequate compensation**
   Development and use of AI in law does not just passively listen to the voices of all individuals and groups that are affected by the operation of 'legal technology'. Developers should actively seeks and involves these voices. Through their labour these communities add value to the eventual product, value and labour that has to be adequately compensated and acknowledged.
10. **Principle of diversity and representativeness**
    The groups and individuals who through their labour and expertise inform the development of 'legal technology', also and in particular through ethical advisory boards and other governance structures, should be representative of the society that the technology will serve and reflect its diversity.
11. **Principle of non-discrimination**
    Biased and discriminatory practices are incompatible with the rule of law ideal and its promise of justice for all. Developers and operators of 'legal technology' must demonstrably ensure through the choice of proper design methodologies (e.g. vetting of input data etc), choice of technology (algorithms that are interpretable and/or have been debiased), testing (both during design and after deployment), and an analysis of other forms of disparate impact on communities, informed by their lived experience that they do not unjustly discriminate against individuals and communities.
12. **Principle of temporal contestability**
    All 'legal technology' projects should anticipate that laws change in response to changing societal norms. The deployment of 'legal technology' must not lock in or constrain future legislators, and must permit challenges and changes to current laws through the democratic process or through judicial review.

Principles for Responsible Use of AI in Law
Accountability: Principles for Responsible Development and use of AI in the Public Sector

13. **Principle of continuous empirical validation**
    Ethical use of AI in law requires rigorous and realistic testing both pre- and post deployment, with robust methods to share findings about problems between users, and users and regulatory agencies.

14. Principle of contestability

The use of AI must not limit the right to contestation of a decision where the legal system provides such a right. Contestation requires typically that appropriate reasons for a decision are given.

Principles of justification and explanation

15. Principle of adequate justification

For applications of AI in law to be ethically defensible it is not sufficient that 'an' explanation of the decision is given. Rather, to be valid the explanation has to respect the internal logic of the legal system in question, and give appropriate reasons that match the duties to justify decisions that we currently impose on human decision makers such as judges or public administrators.

16. Principle of sufficient explanation

AI in law must be explainable enough to allow adequate allocation of responsibility and of the duty to provide remedies in those cases where AI applications may have caused unjustified harm. Where, exceptionally, such an explanation is not possible, comprehensive no-fault compensation or similar safeguards are necessary.

17. Principle of practical and effective redress

Any evaluation of the ethical soundness of a 'legal technology' project needs to identify the remedies that will be provided to challenge decisions, to ensure that incorrect decisions can be detected, making sure that these remedies are adequate, and that those harmed are given appropriate information and support to exercise their rights effectively.

Justiciability: Principles for Responsible Use of AI by the Judiciary

18. Principle of public justice

Public justice, notably the publicness of the trial, is a cornerstone of the fair trial. The use of 'legal technology' is ethically harmful if it increases the obstacles for the public to get involved with, and learn about, the justice system. Responsible use of legal technology, by contrast, aims to lower these obstacles, while preserving legitimate privacy interests of those involved in the proceedings.

19. Principle of equal access to justice

The use of 'Legal technology' should strive to maximize access to legal sources for all. It should in particular be used to widen access to groups that have historically faced significant access barriers. It must not lead to new obstacles, including technological and economical obstacles. Design for accessibility that considers a wide range of disabling factors, in close consultation with the affected communities, is central for ethical 'legal technology'. Alternative modes of access have to be preserved for those who cannot be accommodated this way.

Responsibility: Principles of Responsible use of AI in Legal Practice

20. Principle of equivalent application of professional ethics rules

Where 'legal technology' carries out operations traditionally carried out by lawyers as members of a regulated profession, the protection for the affected citizen should not allow any circumvention of the core principles that govern professional conduct of practicing lawyers, notably (but not limited to) the norms set out by the International Bar Association concerning: independence, honesty, integrity and fairness, conflicts of interest, confidentiality and professional secrecy, clients interests, lawyers' undertaking, clients' freedom, property of clients and third parties, competence and fees.

21. Principle of ultimate responsibility

The choice to use or not to use a given AI as a tool is an exercise of professional judgement. As members of a profession, lawyers have particular ethical responsibility for the tools they chose for the discharge of their responsibilities.

22. Principle of collective responsibility

The safe and ethically sound use of AI in the legal domain is a collective responsibility of lawyers and their professional bodies. This means a duty to share experience especially of problems and errors encountered.

23. Principle of technological neutrality

If lawyers or government officials employ 'legal technology' to provide a legal service or to make legally relevant decisions, their duties as professional lawyers remain unqualified, and do not shift to the system developers. The use of technologies does not alter what clients or public may expect of legal professionals.

Bibliography

The references are a curated listing of the most relevant state of the art literature on the ethical dimensions of 'legal technologies', with no claim to being comprehensive, along with some other works referred to in the report.

Allhutter D and others, 'Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective', (2020) 3:5 Frontiers in Big Data

Ashley, KD, Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age (Cambridge University Press 2017)

Baker JJ, 'Beyond the Information Age: The Duty of Technology Competence in the Algorithmic Society' (2017) 69 South Carolina Law Review 557

Bankowski Z, White I, and Hahn U (eds), Informatics and the Foundations of Legal Reasoning (Springer 2013)

Barocas, S and Selbst AD, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671

Bayles MD, Professional Ethics (Wadsworth 1981)

Bench-Capon T and others, 'A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law' (2012) 20 Artificial Intelligence and Law 215

Brownsword R, 'Technological Management and the Rule of Law' (2016) 8 Law, Innovation and Technology 100

Burk DL, 'Algorithmic Legal Metrics' (2021) 96 Notre Dame Law Review 1147

Campbell RW, 'Artificial Intelligence in the Courtroom: The Delivery of Justice in the Age of Machine Learning' (2020) 18 Colorado Technology Law Journal 323

Chishti S and others, The Legaltech Book: The Legal Technology Handbook for Investors, Entrepreneurs and Fintech Visionaries (Wiley 2020)

Cieslak M, 'Virtual reality to aid Auschwitz war trials of concentration camp guards' (BBC News, 20 November 2016) <https://www.bbc.co.uk/news/technology-38026007> accessed 22 February 2021

Citron DK, 'Technological Due Process' (2008) 85 Washington University Law Review 1249

Compagnucci, MC and others, Legal Tech and the New Sharing Economy (Springer 2019)

Custis T and others, 'Westlaw Edge AI Features Demo: KeyCite Overruling Risk, Litigation Analytics, and WestSearch Plus', in Floris Bex and others, Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL '19 (Association for Computing Machinery 2019)

Deakin S and Markou C (eds), Is Law Computable?: Critical Perspectives on Law and Artificial Intelligence (Hart Publishing 2020)

Dignum V, Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way (Springer 2019)

Diver L, Digisprudence: The Design of Legitimate Code (2021) 13 Law, Innovation and Technology forthcoming

Dubois C, 'How Do Lawyers Engineer and Develop LegalTech Projects? A Story of Opportunities, Platforms, Creative Rationalities, and Strategies', (2020) 2 Law, Technology and Humans

Duff A, and Duff RA, Punishment, Communication, and Community (Oxford University Press 2001)

Dworkin R, Taking Rights Seriously (A&C Black 2013)

Fuller L, The Morality of Law (Yale University Press 1969)

Greenleaf G, Chung P and Mowbray A, 'Building Datalex Decision Support Systems: A Tutorial on Rule-Based Reasoning in Law' (2017) UNSW Law Research Paper No. 17-68

Hartung M, Bues MM and Halbleib G (eds), Legal Tech: A Practitioner's Guide (Hart/Nomos 2018)

Hildebrandt M and Tielemans L, 'Data protection by design and technology neutral law' (2013) 29 Computer Law & Security Review 509

Hildebrandt M, 'Data-Driven Prediction of Judgment. Law's New Mode of Existence?' Forthcoming in Collected Courses of the Academy of European Law (Oxford University Press)

Hart HLA, The Concept of Law (Oxford University Press 1961)

Hoffmann-Riem W, 'Legal Technology/Computational Law' (2021) 1 Journal of Cross-Disciplinary Research in Computational Law

Hoogland J and Henk J, 'Professional Autonomy and the Normative Structure of Medical Practice' (2000) 21 Theoretical Medicine and Bioethics 457

Jonas H, The Imperative of Responsibility: In Search of an Ethics for the Technological Age (University of Chicago Press 1985)

Kennedy R, 'E-Regulation and the Rule of Law: Smart Government, Institutional Information Infrastructures, and Fundamental Values' (2016) 21 Information Polity 77

Law Society, Lawtech: A Comparative Analysis of Legal Technology in the UK and in Other Jurisdictions (The Law Society 2019)

Law Society, Technology, Access to Justice and the Rule of Law: Is Technology the Key to Unlocking Access to Justice Innovation? (The Law Society 2019)

Livermore MA and Rockmore DN (eds), Law as Data: Computation, Text, & the Future of Legal Analysis (Santa Fe Press 2019)

Mohun J and Roberts A, 'Cracking the Code: Rulemaking for Humans and Machines' (OECD 2020)

Morison J and Harkens A, 'Re-Engineering Justice? Robot Judges, Computerised Courts and (Semi) Automated Legal Decision-Making' (2019) 39 Legal Studies 618

Pasquale F, New Laws of Robotics: Defending Human Expertise in the Age of AI (Harvard University Press 2020)

Pistor K, The Code of Capital: How the Law Creates Wealth and Inequality (Princeton University Press 2019)

Radbruch G, 'Legal Philosophy', in Kurt Wilk (tr and ed), The Legal Philosophies of Lask, Radbruch, and Dabin (Harvard University Press 2014)

Spiekermann-Hoff S, Digitale Ethik: Ein Wertesystem für das 21. Jahrhundert (Droemer 2019)

Surden H, 'Machine Learning and Law' (2014) 89 Washington Law Review 87

Susskind R, The Future of Law (Oxford University Press 1996)

Tomlinson J, 'Justice in Automated Administration', (2020) 40 Oxford Journal of Legal Studies 708

Waldron J, 'The Rule of Law and the Importance of Procedure', (2011) 50 Nomos 3

Whalen R, Computational Legal Studies: The Promise and Challenge of Data-Driven Research (Edward Elgar 2020)

Wyner A and Casini G (eds), Legal Knowledge and Information Systems JURIX 2017: The Thirtieth Annual Conference (IOS Press 2017)

Yeung K, ‘Algorithmic Regulation: A Critical Interrogation’ (2018) 12 Regulation & Governance 505

Zalnieriute M and Bell F, 'Technology and the Judicial Role' in Gabrielle Appleby and Andrew Lynch (eds), The Judge, the Judiciary and the Court: Individual, Collegial and Institutional Judicial Dynamics in Australia (Cambridge University Press 2020)

Zeleznikow J, ‘Building Decision Support Systems in Discretionary Legal Domains’ (2000) 14 International Review of Law, Computers & Technology 341

7

MEDIA & TECHNOLOGY

AI in Media & Technology Sector: Opportunities, Risks, Requirements and Recommendationss

Authors

Jo Pierson
Professor of Media, Innovation and Technology at Vrije Universiteit Brussel, Belgium

Stephen Cory Robinson
Senior Lecturer/Assistant Professor in Communication Design at Linköping University, Norrkoping, Sweden

Paula Boddington
Senior Research Fellow, New College of the Humanities London, UK

Patrice Chazerand
Director at DIGITALEUROPE

Aphra Kerr
Professor of Sociology at Maynooth University and Maynooth lead of the ADAPT Centre for Digital Media Technology, Ireland

Stefania Milan
Associate Professor of New Media and Digital Culture, University of Amsterdam

Fons Verbeek
Full Professor in Bio-Imaging and Bio-Informatics, Leiden Insitute of Advanced Computer Science

Cornelia Kutterer
Senior Director, Rule of Law & Responsible Tech, European Government Affairs at Microsoft

Evdoxia Nerantzi
European Government Affairs at Microsoft

Elizabeth Crossick
Head of Government Relations at RELX

ABSTRACT

As AI systems increasingly pervade modern society and lead to manifold and diverse consequences, the development of internationally recognized and industry-specific frameworks focusing on legal and ethical principles is crucial. This report aims at (a) understanding how the 7 Key Requirements for Trustworthy AI impact the Media and Technology sector (MTS) and at (b) putting forward guidelines to ensure compliance with the 7 Key Requirements.

The report identifies four application areas of AI MTS, i.e. automating data capture and processing, automating content generation, automating content mediation and automating communication. Subsequently, the 7 Key Requirements are discussed within each of the four identified themes. Ultimately, recommendations are made to ensure that AI development and adoption in Media and Technology sector is compliant with the 7 Key Requirements. Three clusters of recommendations are proposed: (1) addressing data power and positive obligations, (2) empowerment by design and risk assessments and (3) cooperative responsibility and stakeholder engagements.

Keywords:

Artificial Intelligence, Media and Technology Sector, Trustworthy AI

1. Introduction

AI systems are increasingly pervasive in the individual, organisational, and institutional layers of modern society. Laying the foundations for a "Good AI Society", the multi-stakeholder initiative AI4People initiated the development of internationally recognized and industry-specific frameworks, considering ethics principles.[1] [2] This report examines the large-scale deployment of AI (understood as intelligent and/or autonomous systems) in the Media and Technology sector (MTS). Within this sector, the report lays out four central themes: *automating data capture and processing, automating content generation, automating content mediation*, and *automating communication*. For each of these themes, the report identifies overarching opportunities and risks stemming from the use of AI.

In this report, the Media and Technology Committee – chaired by Jo Pierson – puts forward guidelines to ensure compliance with the 7 Key Requirements for Trustworthy AI. The 7 Key Requirements for Trustworthy AI were originally developed by the European Commission's High-Level Expert Group on Artificial Intelligence[3] and include:

1. Human agency and oversight: Allowing humans to make informed decisions and ensuring human oversight mechanisms;
2. Technical robustness and safety: Ensuring resilient and secure AI systems, a fall back plan, accuracy, reliability and reproducibility;
3. Privacy and data governance: Respecting privacy and ensuring protection, governance, quality of and access to data;
4. Transparency: Ensuring transparent, explainable, and traceable AI models;
5. Diversity, non-discrimination and fairness: Ensuring accessibility to all while diminishing prejudice, discrimination, and unfair bias;
6. Societal and environmental well-being: Ensuring sustainable and environmentally friendly AI systems and considering social and societal impact;
7. Accountability: Ensuring responsibility and accountability of AI systems and their outcomes and adequate redress.

This report is structured in three main sections. After the introduction, Section 2 delineates the Media and Technology sector based on Garnham's framework of mediation and identifies the application areas of AI. This section also provides the

1 All co-authors of this paper constitute the AI4People-Automotive Committee. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayen, E. (2018). AI4People: An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines 28, 689–707.

2 Hagendorff, T. (2020). The ethics of AI ethics. An evaluation of guidelines. Minds and Machines, 30: 99–120.

3 HLEG (2019). Ethics guidelines for trustworthy AI. Brussels: Independent High-Level Expert Group on Artificial Intelligence.

definition of AI used throughout this report. Section 3 summarizes the state-of-the-art in European AI governance. Section 4 is divided into two parts. First, the 7 Key Requirements are discussed within the four identified themes of the Media and Technology sector. Second, recommendations are made to ensure that AI development and adoption is compliant with the 7 Key Requirements in the Media and Technology sector.

2.
Conceptual framework for AI in Media and Technology Sector

a. Definition Media and Technology Sector

This section delineates the Media and Technology sector and identifies the application areas of AI. This is no simple matter given the broad field and fast evolution of the Media and Technology sector due to constant innovation. To begin with, MTS involves every form of technologically supported interaction and communication within an ecosystem where they intersect with specific dynamics, i.e. personalization algorithms. This refers to digital media, i.e. digitised traditional content media, and digital platforms which act as socio-technological intermediating architectures and infrastructures enabling and steering interaction and communication between users through collection and circulation of data.[4] These data are collected, processed and used in MTS for many purposes, among which automated personalisation of (recommendations for) content (e.g. news) and advertising (e.g. targeted advertisements). We observe how especially apps are taking an increasingly prominent place in the MTS, as a large amount of digital media communication today happens via apps, while being embedded within a wider ecosystem.

To determine the essential dimensions of the MTS and to situate AI, we adopt the three main components of Garnham's concept of mediation.[5]

- The first dimension includes **human agents** (*human intermediaries*) which refer to people themselves being mediators, e.g. 'gate-keepers' in (citizen) journalism and news production.
- The second dimension includes **content** (*systems of symbolic representation*) in the form of language and symbols, i.e. how humans produce ('encode') text and consume ('decode') text and what happens to the meaning when it is transported and mediated through languages and cultures.
- The third dimension, which includes **technological systems** (*technological tools in media systems*), prevails when it comes to AI applications in the MTS. The dimension refers to the role and meaning of media systems and related technologies.

4 Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. The information society, 34(1), 1-14.
5 Garnham, N. (2000). Emancipation, the media, and modernity: arguments about the media and social theory. New York: Oxford University Press.

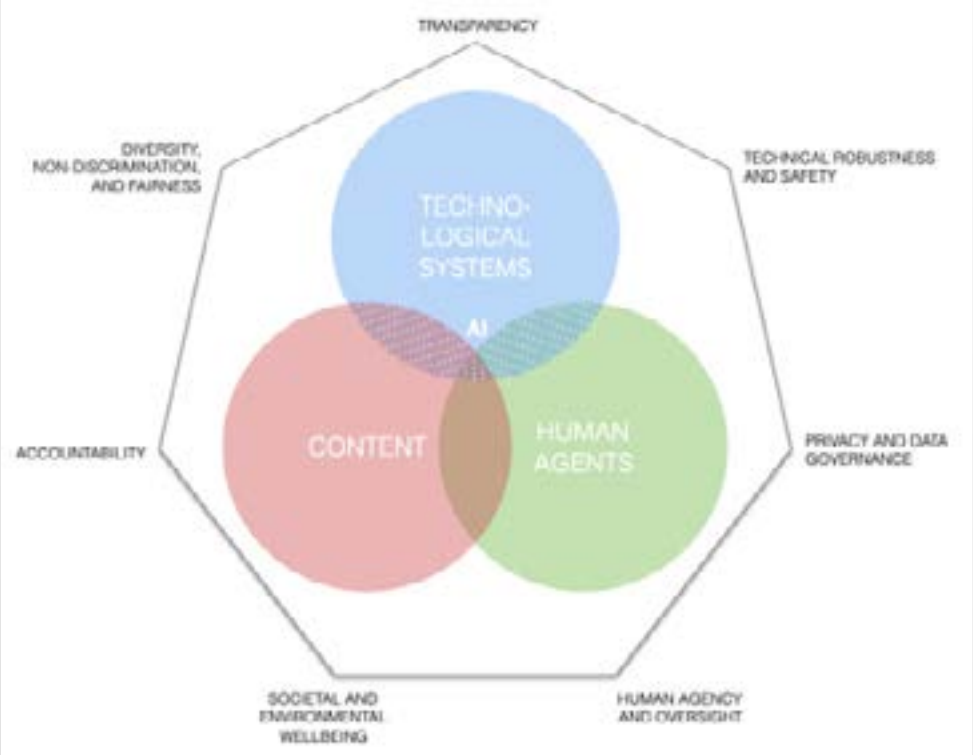


Figure 1: 'Trustworthy AI' heptagon in the Media and Technology sector (own figure)

Within the proposed frame, we identify the following examples of application areas of AI in the MTS which this report examines in the light of the 7 Key Requirements of 'Trustworthy AI':

Human agents +Technological systems	• Automating tools for journalists (Twitter analysis) • Data journalism tools • Newsletter and Customer Relationship Management (CRM) tools • Social media advertising • Etc.
Content + Technological systems	• Computational journalism and robot journalism » Augmenting journalistic practices (high level of AI autonomy): information sharing and gathering, content generation (e.g. sports and financial reporting), revision and distribution • Chatbots, cobots, robots • Deepfakes production and diffusion based on AI • Search engines (algorithm-based technologies) • Smart speakers, voice assistants, new forms of communication (VR/AR) » Speech and face recognition systems » Image analysis software • Marketing automation, programmatic advertising (real-time bidding) and online behavioural advertising • Etc.
Human agents + Content + Technological systems	• Digital media » Journalistic practices (low or medial level of AI autonomy): information sharing and gathering, content generation (e.g. sports reporting), revision and distribution • Digital intermediaries » Digital platforms • General-purpose social media platforms, e.g. Facebook, Twitter • Specific-purpose platforms, e.g. Craigslist, Upwork • News via social media by journalists, citizen journalists and people • Messaging services • Personalisation algorithms (e.g. based on inferential predictive analytics) • Video games • Etc.

Table 1: AI application areas in the MTS

We see the MTS is highly relevant and even exemplary for discussing opportunities, risks and requirements for Trustworthy AI. This is related to several factors. The sector is more directly user-facing compared to other sectors such as energy or automotive sector, with, for example, social media platforms being essential for social interaction and information sharing. This means that people might peg the confidence they should have in digital technology to how much they can trust social media platforms. However, at the same time, the MTS offers the opportunity to provide AI with a promising front office, by realistically framing doom stories and possibly showcasing the advantages of cutting-edge technology. In that way, people can learn the ropes of empowerment in an environment which is more familiar, or less forbidding than anything related to health or mobility. Consequently, the MTS lends itself very well for analysing and discussing Key Requirements for Trustworthy AI in Europe.

b. Definition AI

The AI HLEG (2019) defines Artificial Intelligence as human designed systems which are implemented in the digital or physical environment in the form of software-based systems or possibly hardware devices. Being given a certain goal, AI collects data and assesses the information based on reasoned decision-making in order to suggest relevant actions to achieve the goal. This process is guided by a set of symbolic rules or a numeric model as well as the ability of AI to learn from their environment and previous outputs.[6] The COM (2018) on Artificial Intelligence in Europe emphasizes the intelligent and to a certain extent autonomous behaviour of AI systems.[7]

In fact, AI can be defined as machines that acquire cognitive capabilities such as learning, taking decisions, communicating and interacting based on digital data. Machine learning (ML) and algorithms are two essential features of this process. An algorithm is a software which processes input, i.e. data, based on described rules and selects the relevant information for the user. Moreover, AI is capable of prediction-making, decision-making and problem-solving.[8]

6 HLEG (2019). A Definition of AI: Main Capabilities and Disciplines. Brussels: Independent High-Level Expert Group on Artificial Intelligence.
7 COM (2018). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Artificial Intelligence for Europe. European Commission, 237 final.
8 Just, N., & Latzer, M. (2017). Governance by algorithms: reality construction by algorithmic selection on the Internet. Media, culture & society, 39(2), 238-258.

Examining the level of autonomy of AI, Boucher (2019) differentiates between two waves in AI development.[9] The first wave, called symbolic artificial intelligence, is rather human-centered, which means that even though the AI system performs tasks autonomously, the decision-making process is still guided by humans (*human in the loop*). The system's intelligence stems from the encoding of human expertise and, hence, makes the process and output more comprehensible for humans. In the second wave, called data-driven machine learning, algorithms gain more autonomy and become rather independent from human expertise as they train themselves from data and statistics (from *human-over-the-loop* to *human-out-of-the-loop*). Striking a balance between data-driven and human-centred expertise and assistance is important especially within the scope of the MTS sector as automatisation processes increasingly penetrate journalism and communication activities, a core feature of European democratic processes.

Given that AI systems make recommendations and provide normative solutions, the notion of trust is important to be examined.[10] According to the AI HLEG, trustworthiness should represent a "prerequisite for people and society to develop, deploy and use AI".[11] Hence, the MTS needs to be continuously vigilant that AI systems stay trustworthy even after having been developed, implemented and/or used. People should not be "nudged" or forced to use systems they do not trust or that do not adhere to the 7 Key Requirements. In this light, the next section briefly examines how this has been tackled by the EU to date and what this in particular means for AI applications in the MTS.

9 Boucher, P. (2019). How artificial intelligence works. Brussels: European Parliament Research Service.ù
10 Ferrario, A., Loi, M., & Viganò, E. (2019). In AI We Trust Incrementally: A Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. Philosophy & Technology. doi: 10.1007/s13347-019-00378-3
11 HLEG (2019). Ethics guidelines for trustworthy AI. Brussels: Independent High-Level Expert Group on Artificial Intelligence.

3. European AI Governance for MTS

The EU's long-standing emphasis on democratic values and the rule of law also shapes its technology governance approach. This is especially relevant for the MTS where automated-decision making technologies and algorithm-dependent processes are becoming increasingly essential. We take a closer look at how AI governance is taking form in the EU, with a focus on MTS-related issues.
For the purpose of this report the High-Level Expert Group on Artificial Intelligence (AI HLEG)[12] was an important initiative, being appointed by the European Commission in June 2018. Since then, the AI HLEG's work has been considered as substantial in defining a "European" governance approach centred around the concepts of "ethical" and "trustworthy" AI. The AI HLEG bases its considerations on three key requirements for AI: legal (i.e. AI should comply with the law); ethical (i.e. AI should fulfil ethical principles); and robust (i.e. AI should be built safely and on the highest quality standards). In July 2020, the AI HLEG published their final Assessment List for Trustworthy Artificial Intelligence (ALTAI) for all relevant stakeholders, particularly those involved in developing and deploying AI systems, to self-assess compliance of specific AI use cases with the 7 Key Requirements for Trustworthy AI.

The European Commission (EC) also incorporated the AI HLEG recommendations in their latest White Paper on Artificial Intelligence.[13] The document sets forth to promote and develop AI based on European values, following a regulatory and investment-based approach. The EC refers to the MTS particularly in the context of protecting fundamental human rights and ensuring legal certainty.[14]

Specifically, the EC highlights the use and potential impact of AI (1) for information selection and content moderation by online intermediaries; (2) in tracing people's daily habits; and (3) in creating information asymmetries by which citizens might be left powerless. The EC is particularly concerned about some potential AI systems' features, such as "opacity ('black box-effect'), complexity, unpredictability and partially autonomous behaviour",[15] in overseeing and enforcing the existing EU legal fundamental rights framework. This may be the reason for the introduction of specific rules for 'high-risk' AI systems in a possible forthcoming EU regulatory framework for

12 European Commission. (2019). High Level Expert Group on Artificial Intelligence. Retrieved on May 20, 2020, from https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
13 European Commission. (2020). White Paper on Artificial Intelligence. A European approach to excellence and trust. Retrieved on March 20, 2020, from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
14 European Commission. (2020). White Paper on Artificial Intelligence. A European approach to excellence and trust. Retrieved on March 20, 2020, from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
15 European Commission. (2020). White Paper on Artificial Intelligence. A European approach to excellence and trust. Retrieved on March 20, 2020, from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf, 12

AI systems. An AI system could be considered as 'high-risk' if "both the sector and the intended use involve significant risks"[16], particularly if safety, consumer rights or fundamental rights are at stake. If an AI system meets the 'high-risk' criteria, compliance with strict requirements and oversight would be mandatory. The EC explicitly sets out the protection of the following EU rights:

- Fundamental rights: Free expression; political freedoms; personal data protection; privacy protection; non-discrimination.
- Legal certainty: Safety; liability; cybersecurity.

These fundamental EU rights are relevant for the MTS since information, communication and mediation activities are all intrinsically linked and somewhat a prerequisite for democracy and the rule of law in the EU. Particularly relevant for the MTS is that the White Paper specifically mentions "online intermediaries" and their responsibility in adequately safeguarding the abovementioned rights as required by EU legislation. Further, the EC underlines that citizens should clearly be aware about their interactions "with an AI system, and not a human being."[17] According to the specific context in which the AI application operates, the EC emphasises "objective, concise and easily understandable" information provision. Next to the White Paper on Artificial Intelligence, the European Commission provides an interpretation of the existing safety and liability framework specifically for Artificial Intelligence, the Internet of Things and robotics.[18] Further, the documents apply in addition to key requirements for protecting data subjects and their data, as set out by the EU data protection legislation (GDPR).

In October 2020, the European Parliament released two legislative initiatives to develop an ethics framework for AI and a civil-oriented liability framework for AI causing damage. The first initiative calls for a legal framework outlining the ethical principles and legal obligations for AI following guiding principles such as human-centric and human-made AI, safety, transparency and accountability, safeguards against bias and discrimination, right to redress, social and environmental responsibility, and respect for privacy and data protection.[19] The second initiative encourages the development of a civil-oriented liability framework, calling for liability of humans

16 European Commission. (2020). White Paper on Artificial Intelligence. A European approach to excellence and trust. Retrieved on March 20, 2020, from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf, 17

17 European Commission. (2020). White Paper on Artificial Intelligence. A European approach to excellence and trust. Retrieved on March 20, 2020, from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf, 20

18 European Commission. (2020). Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics. Retrieved on March 20, 2020, from https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1593079180383&uri=CELEX%3A52020DC0064.

19 García del Blanco, I. (2020). Report with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies. Retrieved on November 1, 2020, from https://www.europarl.europa.eu/doceo/document/A-9-2020-0186_EN.html.

when operating with high-risk AI activity.[20] Moreover, the European Parliament published a report on intellectual property rights. The report urges to distinguish between AI-assisted human creations and AI-generated creations and encourages an effective intellectual property rights system (IPR) as well as safeguards for the EU's patent system to protect innovative developers.[21]

Considering the indicated European AI policy initiatives, this report contributes to the EU AI governance process by establishing an ethical framework for AI applications in the MTS.

4.
Research questions

The Media and Technology committee consists of 12 members representing different stakeholders of academia and media and technology industry. The goal is to establish a concerted perspective on the meaning and significance of the HLEG 7 Key Requirements for Trustworthy AI in relation to the MTS. For this, the committee held regular gatherings to discuss implications of the requirements in their respective expertise and industry. The multistakeholder procedure for developing the main research questions, consecutive outcomes and the final report was organised as follows:

- Discussing 7 Key Requirements and specific cases based on committee members' expertise and practical experience;
- Asking members to submit case studies containing best and worst practice use cases of AI in the MTS;
- Members submitted their cases;
- Discussion of the submitted cases;
- Members provided additional information, literature and explanation on cases;
- Committee Chair and Advisors scanned all cases for keywords and overlapping issues and best/worst practices;
- Committee Chair and Advisors grouped submitted AI case studies into four categories (themes);
- Committee Chair and Advisors cross-combined four themes with 7 Key Requirements;

20 Voss, A. (2020). Report with recommendations to the Commission on a civil liability regime for artificial intelligence. Retrieved on November 1, 2020, from https://www.europarl.europa.eu/doceo/document/A-9-2020-0178_EN.html.
21 Séjourné, S. (2020). Report on intellectual property rights for the development of artificial intelligence technologies. Retrieved on November 01, 2020, from https://www.europarl.europa.eu/doceo/document/A-9-2020-0176_EN.html.

- Committee Chair and Advisors identified tensions within the four themes;
- Cross-combination of four themes with 7 Key Requirements was first discussed in-depth, after which the view of the committee was further validated and visualised through an online form and interviews among members;
- Members proposed recommendations to ensure compliance with the 7 Key Requirements within the four main themes of the MTS;
- Committee Chair and Advisors identified three prevailing recommendation clusters.

I.
How do the 7 Key Requirements impact the Media and Technology sector?

AI technologies are used in various MTS areas and for various purposes. Given the related manifold and diverse consequences of AI, the analysis and discussion of the 7 Key Requirements for trustworthy AI in MTS is structured according to four main *MTS AI application and use themes*:

a) Automating data capture and processing;
b) Automating content generation;
c) Automating content mediation;
d) Automating communication.

The four themes are mapped in line with the typical (big) data life cycle of data capture, processing and interpretation, preparation and creation, and usage.[22] The four MTS AI application and use themes aim (1) to be mutually inclusive and (2) to largely capture all relevant cases which fall under the MTS in the scope of this report. This is made tangible in the following paragraphs by focusing on concrete examples, when discussing the Key Requirements for each individual theme. This approach allows for a content-based discussion instead of discussing various cases and impacts under each key requirement. This bottom-up, practically oriented methodology also allows to discuss tensions between the 7 Key Requirements.

22 Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. Communications of the ACM, 57(7), 86-94.

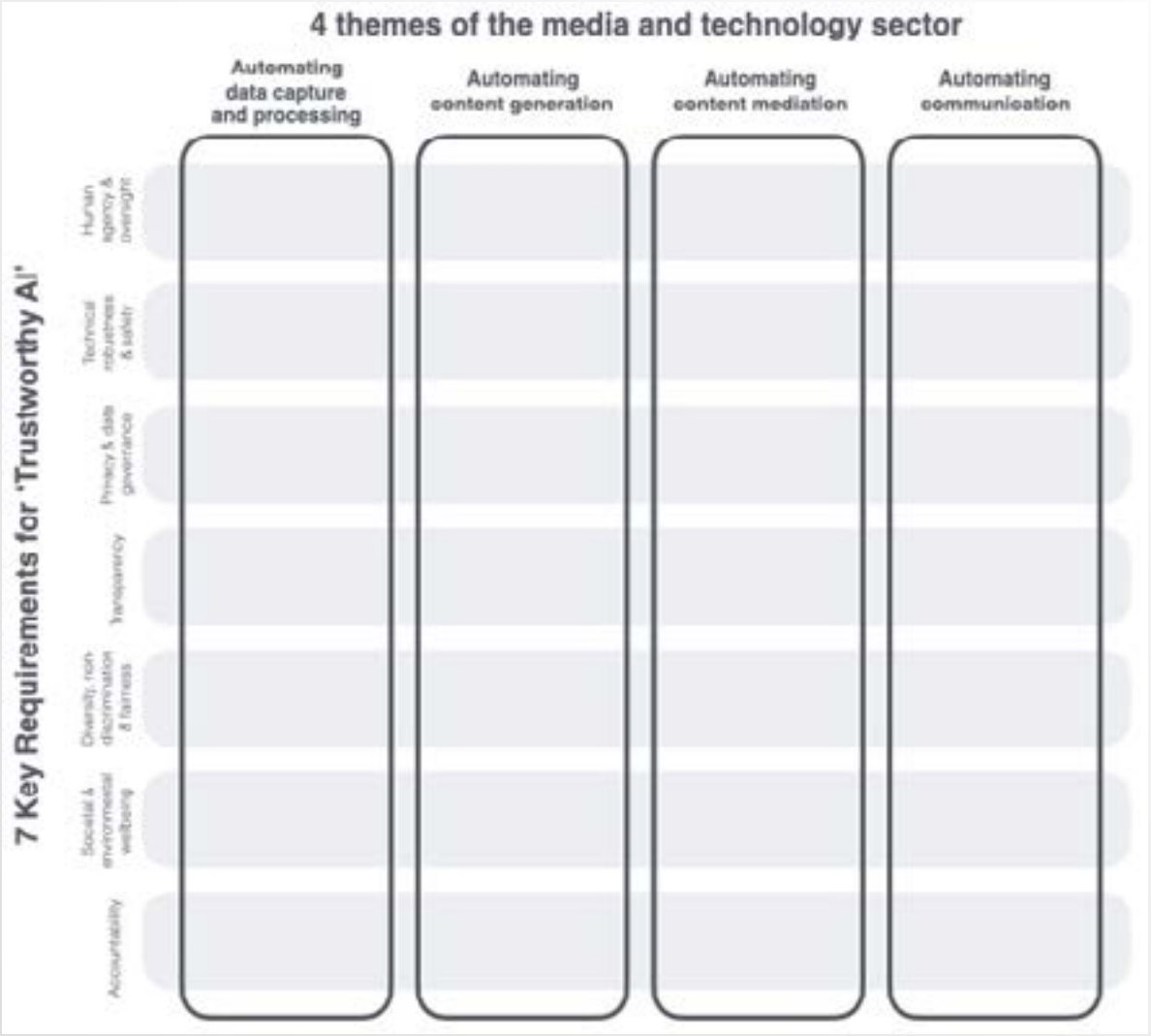


Figure 2: The 7 Key Requirements for "Trustworthy AI" in relation to the four themes of the Media and Technology Sector.

a. Theme 1: Automating data capture and processing

The first theme, **automating data capture and processing**, entails a variety of AI technologies concerned with the systematic capture and processing of data in the MTS. This typically includes data capture and processing by digital media, platforms and websites for reasons of personalisation, profiling, inferential predictive analytics, targeted advertising, etc. However, this type of automation also includes emotional AI in the form of facial and voice recognition systems as well as GPS/location tracking,

contact tracing apps, and VR/AR headsets.[23] Within the 'Trustworthy AI' heptagon (Fig. 1) this theme is concerned with the diminishing aspect of humans as agents against augmenting impacts of technological systems on content.

The principles *human agency and oversight* as well as *privacy and data governance* have a high impact on automated data capture and processing. First, the EU's legislative framework, in particular data and consumer protection standards, protects individuals' fundamental rights to make informed and independent choices. This also applies in relation to automated data capture and processing AI systems: *human agency and oversight* as well as *privacy and data governance* demand that citizens should always be able to decide if and how they choose to use a certain service or be inadvertently tracked by it. In case of using a MTS service, there is the right to decide on what and how much data will be collected, what it would be used for, where it would originate from, and how it would be shared. Given the advertising-driven business model for a significant part of MTS, special attention is needed on how data capturing and processing takes shape with regard to adtech and marketing automation. This is particularly relevant for online behavioural advertising (OBA), where internet users' behavioural data (website visits, clicks, mouse movements, etc.) and metadata (browser type, location, IP address, etc.) are collected and processed to create profiles used to personalise ads and to improve conversion rates. Recent events have shown that especially automated advertising systems of real-time bidding (RTB) have been capturing and processing in possibly prohibited and unethical ways.[24] RTB in ad auctions is the system by which advertisers bid on the possibility of instant targeted advertising to website visitors by using personal data that is collected through tracking and is shared with all bidders. Even advertisers who do not win the auction receive personal data in order to ascertain their interest in the auction. Some advertisers are reported to participate in the auctions merely to enrich their data sets. The targeting is based on profiles of users built via the extensive and persistent tracking of online and possibly offline activities (e.g. via cookies or pixels). The profiles contain categories of users' past behaviour, but also inferred preferences and affinities, being often sensitive categories protected by the GDPR. For example, Google and several data brokers have been accused of violating EU's data protection rules by harvesting and processing people's personal data to build detailed online profiles, including information on sexual orientation, health status and religious beliefs.[25] Additionally, the Norwegian consumer council investigated the data traffic from popular mobile apps. This revealed a number

23 For example immersive mixed reality headset (i.e. Microsoft HoloLens).
24 Information Commissioner's Office (2019). Update report into adtech and real time bidding, 20 June 2019. Retrieved on November 13, 2020, from https://ico.org.uk/media/about-the-ico/documents/2615156/adtech-real-time-bidding-report-201906.pdf
25 Scott, M., Manacourt, V. (2020). Google and data brokers accused of illegally collecting people's data: report. in: POLITICO, 21 September 2020. Retrieved on November 12, 2020, from https://www.politico.eu/article/google-and-data-brokers-accused-of-illegally-collecting-data-report/amp/

of serious privacy infringements and a large amount of illegal data sharing and processing.[26] Academics and data protection practitioners have made proposals to address these type of privacy infringements. Wachter and Mittelstadt suggest introducing the "right to reasonable inferences" by which meaningful control and choice over inferences and profiles are granted to data subjects.[27] This would be particularly relevant for high-risk inferences that are privacy invasive or reputation damaging and have low verifiability in the sense of being predictive or opinion-based.[28] Envisaged as an ex-ante mechanism to provide justification for the reasonability of an inference, disclosing relevance of the data in question, relevance of the inferences drawn, accuracy and statistical reliability of the methods used, these disclosures should be accompanied by an ex-post mechanism enabling inferences to be challenged. This right should close the gap both of explainability and accountability.

In addition, given the importance of being able to collect and process as much as possible (personal) data for optimising personalisation of content and advertising, special attention is needed to safeguard a level playing field in MTS. Although all players in the media and advertising ecosystem are affected by the GDPR, larger players may be more resilient to regulatory interventions. In case smaller competitors drop away, the consolidation of personal data in fewer hands might also increase, and perversely, negatively affect people's rights and freedoms overall. For that reason, various initiatives have been taken, especially in smaller media markets, to pool data and to process them for the benefit of different (competitive) companies at once.[29]

Automated data capture and processing also takes place in other types of applications (in work, health, leisure time) as well as devices (VR/AR headsets[30]). Especially if emotion-reading and -inferring AI systems were to be adapted on a large scale for partially abled people, the option to not use or be subjected to such AI systems should always be available for the person. Moreover, individuals should be aware if they are being systematically tracked, such as by websites, platforms, apps and cameras, and for which purpose, based on an opt-in regime in line with the GDPR. However,

26 ForbrukerRådet. (2020). Out of control: How consumers are exploited by the online advertising industry. Report by the Norwegian Consumer Council.
27 Pop Stefanija, A. (2019, July 7-11). Algorithmic selfie: on the right to assess algorithmic identity and exercise right of access". Madrid, Spain: IAMCR 2019 Conference.
28 Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. Colum. Bus. L. Rev., 494.
29 van Zeeland, I., Ranaivoson, H., Hendrickx, J., Pierson, J., Van den Broeck, W. & van der Bank, J. (2019). Salvaging European media diversity while protecting personal data. Brussels, Belgium: SMIT Policy Brief #23, Report for Chair 'Data Protection on the Ground' (Media Sector).
30 As such, Microsoft HoloLens - immersive mixed reality headset - can help people who are blind and with low vision learn who is where in their social environment.
Roach, J. (2020). Using AI, people who are blind are able to find familiar faces in a room. Retrieved on May 2, 2020, from https://news.microsoft.com/innovation-stories/project-tokyo/?utm_source=pre-amp

there can also conditions where certain types of capture are required for overriding purposes (e.g. tracking potential terrorists, in case of a substantiated suspicion). The key principle to safeguard *human agency and oversight* also implies that citizens must be aware if their information or face is being recorded, especially if personal data is tracked. Giving their consent in context of AI applications often involves a weighing of benefits and harms of not opting in, resulting in a rather reluctant agreement than genuine willingness. [31] [32]
Given that tracking of data has become an essential part of many platforms and services in MTS, not opting into the conditions often leads to substantial disadvantages for users. A swift implementation of automated systems for data capture and processing during emergency situations may however lead to lower standards of accuracy and ethical oversight (e.g. in the case of COVID-19 contact tracing apps).[33] In addition, the wide adoption of tracking apps may lead to chilling effects in surveillance of free movement and/or individual behaviour.

Moreover, users should have agency over their data when they visit MTS websites or use apps which record certain information on purpose. Users' given consent needs to be purpose-limited and context-specific. The key principle *transparency* would enable increased *human agency and oversight* if AI uses systematic tracking and tracing features. Transparent and explainable automated data capture AI can build human oversight and trust in technologies. However, a challenge to becoming transparent poses the complex nature of AI itself, such as machine learning. It seems to be a difficult matter to agree on what a 'transparent' explanation of the system must contain and to come to an understanding of what sort of information is enough for all types of individuals subject to the data capture and processing. Linking this to consent, the specific requirements of transparency needed to obtain genuine consent could vary from domain to domain.

The principle of *diversity, non-discrimination and fairness* would impact the MTS in a way that algorithmic data capture and processing AI should not discriminate and/or be biased, and promote a stated conception of fairness. Conceptions of what is a fair

31 Apps for coronavirus contact tracing could trace the spread of disease, to understand infection pathways for risk individuals and communities, and could help in delivering resources to where needed. However, the same technology could be used for wider surveillance of populations and for very punitive consequences in some societies, especially in combination with other technology such as facial recognition to monitor citizens' behaviour. Surveillance measures may outlast the need.
Prasso, S. (2020). Corona Virus surveillance helps, but the programs are hard to stop. Retrieved on April 20, 2020, from https://www.bloomberg.com/news/articles/2020-04-06/coronavirus-surveillance-helps-but-the-programs-are-hard-to-stop
32 Gershgorn, D. (2020). We mapped how the Coronavirus is driving new surveillance programs around the world. Retrieved on April 20, 2020 from https://onezero.medium.com/the-pandemic-is-a-trojan-horse-for-surveillance-programs-around-the-world-887fa6f12ec9
33 Van Zeeland, I. & Pierson, J. (2020). Contact tracing apps and solutionism. Position statement for the Future of Privacy Forum's "Privacy & Pandemics: Responsible Uses of Technology and Health Data During Times of Crisis - An International Tech and Data Conference".

distribution of anything in society differs. Subsequently, what a system developer may deem as 'fair' should be explicitly stated, as well as promoted already in the data capture stage (e.g. do not only process data on young people's news preferences if this is further used for providing recommendations to seniors). In relation to automating data capture, the link between *technical robustness and safety* is paramount because automated data capture systems need to fully represent all potential users and other individuals who will be affected by the systems' outcomes, and not only a certain (biased) part of its dataset. In the MTS, those key requirements are particularly relevant for content-related platforms, like search engines (e.g. Google, Bing), social network sites (e.g. Facebook, Twitter) and video sharing services (e.g YouTube, Vimeo). Equally important are to uphold the key requirements *diversity, non-discrimination and fairness* to avoid the simplistic classification of emotions, which could result in unwanted social sorting. More generally, cultural norms in emotions are not yet fully researched, and psychological research would be suited to inform technology developers about the social norms behind public display of emotions. Furthermore, if the AI system could appropriately adopt cultural norms, it would require consideration if reinforcing certain cultural norms is desirable or not. It is, more fundamentally, worth considering which ethically legitimate purposes could be served by processing data on human emotions at scale.

Technical robustness and safety are also highly important to not market any AI systems for which the impact is not well-researched, and which are not yet fully developed, based on the precautionary principle. The risk is to release it too early. As such, the 'misreading' of emotions could create serious damage to both users and corporate reputation. This is also why the auditing of automated and algorithmic data capture and processing AI systems is key: *accountability* enables a more comprehensive assessment of the purpose, development and deployment of automated data capture and processing AI and additionally enhance *transparency*. Finally, *accountability* can further be improved through multi-stakeholder deliberation, maintenance and oversight, where also citizens and civil society organisations are represented in meaningful way.

Prospectively, automating data capture and processing technology could allow more immersive interactions with surrounding environments by means of Virtual Reality (VR) or Augmented Reality (AR) technologies.[34] Large amounts of data could determine the large-scale nudging by "recommendations", as such for online maps, services or products. Automating data capture and processing could enable targeted consumer choice but likewise decrease *human agency and oversight*. More and better datasets by automating data capture and processing could create powerful nudges based

34 Pollock, D. (2019). Digital billboards open-up advertising to blockchain, artificial intelligence, and cryptocurrency. Retrieved on April 20, 2020, form https://www.forbes.com/sites/darrynpollock/2019/04/18/delving-into-digital-advertising-as-blockchain-cryptocurrency-iot-ar-and-ai-enter-the-frame/

on emotional appeals which one is unable to rationally and cognitively process. This is why *human agency and oversight* are key requirements as long as AI technologies for widespread automating data capture and processing progress. At the same time, prediction based on automating data capture and processing requires considering *accountability*. People should be able to know who is providing the information and what database the prediction is based on, such as models of other people or past behaviour. Several principles come together, as *accountability* towards users and supervisory authorities, to enable *effective independent oversight*, requires *transparency* for data sets to determine whether the captured and processed data indeed supports *diversity, non-discrimination, and fairness*.

b. Theme 2: Automating content generation

The second theme, **automating content generation**, refers to online content produced either fully by automated systems or partly in combination with human agents. Examples of common AI uses in content generation are text-based news reporting apps (based on user preferences)[35] and translation tools, and - in a malign way - disinformation and deepfakes on online platforms. The question remains how much of this type of content is fully automated. The automated element is perhaps more prevalent in the diffusion and amplification of the content rather than the production of it. In addition, an emerging AI application area are creative industries, such as the music and games industry, and creative AI/computing.[36] Considering the 'Trustworthy AI' heptagon (Fig. 1), strong links between technological systems and content become evident. In sum, increasing automation in content generation may provoke an imbalance disfavouring the role of human agents in content generation.

As human agents like journalists play a major role in providing trustworthy information, the aspects of *human oversight*, *accountability*, and *technical robustness* are highly important. AI-driven tools are already employed in journalistic content generation, which relates to the principle of human agency and oversight. In data journalism, for instance, AI helps to identify patterns in large datasets. AI-driven tools can suggest titles and photos, help to find a new topic angle, and produce draft versions of articles. Automated systems assist the journalist in writing the story, but the journalist is still the main storyteller.[37] Thus, a high level of editorial input and human oversight remains; at the same time, publishing articles becomes more efficient. The increasing

35 For example Google News, Apple News, Reddit, Digg, and Flipboard.
36 Amato, G., Behrmann, M., Bimbot, F., Caramiaux, B., Falchi, F., Garcia, A., & Koenitz, H. (2019). AI in the media and creative industries. arXiv preprint arXiv:1905.04175.
37 Willens, M. (2019). Forbes is building more AI tools for its reporters. Retrieved on March 4, 2020, from https://digiday.com/media/forbes-built-a-robot-to-pre-write-articles-for-its-contributors/.

pace and efficiency of news production triggered by automation can put pressure on smaller newsrooms which usually do not dispose large datasets and robust AI-systems.[38] In some news genres, especially those that are rather fact-based, the automation in news generation is higher. For instance, specific natural language processing tools can generate sports articles and financial reporting,[39] while recent projects even involve video reporting[40] Higher automation of content generation can eventually lead to transitions in working opportunities and possible job loss, impacting *societal well-being*.[41] [42] In addition, content produced by AI systems is often not flagged as such to the user and this, hence, links to the importance of *transparency*.

Technical robustness of AI-driven tools in content generation is essential to manage large amounts of data. Data journalism requires robust AI systems to analyse data correctly and to extract "relevant" information. A key issue is as the definition of 'relevant' information today. New forms of news/ information coupled with commercial pressures on the internet are shaping what is presented as 'news' and how it is presented, e.g. clickbait (content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page). How and what information AI systems extract can, ultimately, shape how the reader understands the information, e.g. positive or negative attitude toward a subject. This can be linked to the principle of *diversity, non-discrimination, and fairness*.

Another example of the significance of technically robust AI systems is in preserving practices of European cultural and architectural heritage. AI systems are capable of digitising high volumes of information which is stored in physical form in archives and museum;[43] for instance, IVOW's "Culturally Sensitive Deep Learning model" can create captions for photos generated by natural language processing algorithms.[44]

38 Helberger, N., Eskens, S. J., van Drunen, M. Z., Bastian, M. B., & Möller, J. E. (2019). Implications of AI-driven tools in the media for freedom of expression.
39 Peiser, J. (2019). The Rise of the Robot Reporter (Published 2019). Retrieved on November 13, 2020, from https://www.nytimes.com/2019/02/05/business/media/artificial-intelligence-journalism-robots.html
40 Chandler, S. (2020). Reuters uses AI to prototype first ever automated video reports. Retrieved on May 10, 2020, from https://www.forbes.com/sites/simonchandler/2020/02/07/reuters-uses-ai-to-prototype-first-ever-automated-video-reports/#7eb6a99f7a2a
41 Lindén, C.-G., Tuulonen, H. (Eds.) (2019). News Automation. The rewards, risks and realities of 'machine journalism'. Frankfurt: WAN-IFRA. Retrieved November 22, 2020, from http://immersiveautomation.com/wp-content/uploads/2019/06/WAN-IFRA_News_Automation-FINAL.pdf. sws
42 Srnicek, N. (2017). Platform Capitalism. Polity Press.
43 Ibaraki, S. (2019). Artificial Intelligence For Good: Preserving Our Cultural Heritage. Retrieved on March 6, 2020, from https://www.forbes.com/sites/cognitiveworld/2019/03/28/artificial-intelligence-for-good-preserving-our-cultural-heritage/#200a70094e96https://www.forbes.com/sites/cognitiveworld/2019/03/28/artificial-intelligence-for-good-preserving-our-cultural-heritage/#200a70094e96
44 IVOW. (2020).An AI and Storytelling Startup. Retrieved on May 10, 2020, from https://www.ivow.ai

Technical robustness is also highly relevant in producing and translating texts. Automated translation risks replication biases (e.g. stereotypes, gender and racial biases) and errors from training datasets. This can affect the principle of *diversity, non-discrimination, and fairness*. Given the linguistic diversity in the European Union, robust automated translation is highly important. It preserves linguistic and cultural plurality. For example the ADAPT research center[45] in Ireland aims to develop data sets and intelligent models that automatically translate online content for native speakers of low-resource languages, and make important content available to people in their language of choice. Projects have focused on developing resources for Irish, Serbian, Basque, and non-European languages including Hindi. Their approach is to employ both, AI and human, rather than fully automated systems.[46]

On social media platforms, deepfakes generated by AI-driven tools grow in popularity. These formats simulate a speech or an action, usually of a public persona (such as politicians, celebrities and actors), where the generated content does not correspond to reality but reveals striking resemblance. This is highly problematic because such false information is often generated without the knowledge of the individuals in question and viewers may be unaware the video was tampered with. This applies to the principle of *human agency* and *societal wellbeing*. It can foster the spread of contentious content like 'fake news', disinformation, hate speech and harmful content. One of the most popular videos that went viral in 2019 portrays Marc Zuckerberg claiming to conquer the world. A recent study shows that 72 percent of people reading an AI-generated news story thought it was credible.[47] Another example is the Chinese app Zao which allows people to seamlessly swap themselves into famous movie scenes.[48] Generating deepfakes and producing disinformation challenges media integrity. In addition, it can severely harm individuals through inappropriate and false representation as well as harassment, for example by malign actions like revenge-porn, affecting not just public figures, but also regular, common people. Forms of redress to tackle these issues seem to be underrepresented or do not guarantee general accessibility to citizens. Hence, it can also be linked to the principle of *diversity, non-discrimination, and fairness*.

45 Transforming Global Content. (2020). Retrieved on May 5, 2020, from https://www.adaptcentre.ie/research/transforming-global-content/

46 For example, they have developed a high-quality Irish-English system called Tapadóir to translate documents into Irish for the Irish government. From 2021 all European documents will also have to be translated into Irish and much of this will be done using these automated systems supplemented by Irish language native speakers and translators.

47 Leibowicz, C. (2019). On AI & Media Integrity: Insights from the Deepfake Detection Challenge. Retrieved on April 20, 2020, from https://www.partnershiponai.org/on-ai-media-integrity-insights-from-the-deepfake-detection-challenge/

48 Kambhampati, S. (2019). Perception won't be reality, once AI can manipulate what we see. Retrieved on April 20, 2020, from https://thehill.com/opinion/cybersecurity/470826-perception-wont-be-reality-once-ai-can-manipulate-what-we-see

In the music sector, deploying AI-driven tools links to the principles of *human agency*, *societal wellbeing*, and *diversity*, *non-discrimination*, and *fairness*. While AI may be beneficial for musicians as it could enhance music education and composition, it also causes concerns about, for instance, replacing human creativity and removing the personal aspect of music creation. Furthermore, the human agency in question may affect *societal wellbeing* in hampering the development of human talent. This can result in reducing opportunities for live music and can produce a cycle by which music is generated and experienced online and remotely, with an impact on human social life.[49] Ultimately, the music sector does not represent an urgent human demand to be complemented by AI systems, to justify replacing human labour.

c. Theme 3: Automating content mediation

The third theme, **automating content mediation**, involves automated filtering systems in the distribution and moderation of online content and advertising. AI technologies in content distribution occur in the form of recommender systems for entertainment and social media content, online news aggregators, and programmatic advertising (including RTB) which provide user-specific and context-conform content. A further set of AI systems is employed to moderate content to detect and tackle contentious content like fake news, mis- and disinformation[50], and harmful content.[51] Linking this to the three components in the 'Trustworthy AI' heptagon (Fig. 1) reveals that *online content* is increasingly processed by technological systems either fully automated or assisting human agents.

Employing automated filtering systems in online content and advertising mediation tasks requires a careful consideration of the principles *diversity*, *non-discrimination*, and *fairness* and *human agency and oversight*. For example, we observe how years after the initial research into discrimination in online employment ads, higher salary positions are still advertised to predominantly (assumed) male users.[52] As AI technologies somehow occupy the new role of traditional gatekeepers and doing agenda-setting in the online sphere, they can also co-determine what people see or not see as well as what content users can generate online. This could affect freedom of expression, media diversity and plurality of voices.[53] In the case of algorithmic content distribution, it can constrain access to a diversity of information and create 'filter bubbles' leading to 'echo

49 Castro, A. (2019). We've been warned about AI and music for over 50 years, but no one's prepared. Retrieved on May 1, 2020, from https://www.theverge.com/2019/4/17/18299563/ai-algorithm-music-law-copyright-human
50 EPRS (2019). Regulating disinformation with artificial intelligence.
EPRS (2019). Automated tackling of disinformation.
51 Lacoma, T. (2020). League of Legends Survey Reveals Nearly Every Player Has Been Harassed. Retrieved on May 1, 2020 from https://screenrant.com/league-legends-survey-harassment-toxicity-riot-games-everyone/
52 Datta, A., Tschantz, M.C., Datta, A. (2015). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. arXiv. Retrieved on November 12, 2020 from https://arxiv.org/pdf/1408.6491.pdf

chambers', i.e. personalised content. Especially online platform recommender systems tend to magnify hyperactive users' interests and content, while passive users' interests and content become more invisible.[54] [55] Hence, political microtargeting and opinion formation could become subject to (un)intentional algorithmic manipulation. Furthermore, the datasets as well as developed standards about fairness and non-discrimination for algorithmic filtering systems might contain bias and could leverage discrimination and social sorting. With regards to diversity, non-discrimination, and fairness in the media sector, media recommendation algorithms may worsen the position of smaller countries and their cultural values in media creation.

These issues raise the importance of *human agency and oversight* in online content mediation. Algorithmic filtering systems constrain human agency as users are hampered in choosing which content they receive or if they want to be exposed to algorithmic recommendations at all. Furthermore, human oversight is crucial in detecting and tackling disinformation and harmful content, also in relation to programmatic advertising with advertisers being worried about brand safety with their ads being placed besides contentious content on digital platforms. Take, for instance, the 'infodemic' or large circulation of disinformation and misleading ads during the Covid-19 pandemic (e.g. drinking more water would cure an individual from the disease).[57] [58] [59] Especially in the context of health crises, correct information and reliable sources are particularly important and the lack of it can have severe, even fatal consequences. This case reveals the importance of human oversight in fact-checking the content by professionals, such as health advice. Nevertheless, algorithmic filtering systems are required to master the high volume and fast-paced production of online content. When it comes to content moderation, AI systems are crucial assistants for augmenting human agents in their demanding work of evaluating harmful content such as child abuse, racism, and harassment. This has immediate effects on the physical, mental and *societal well-being* of human content moderators. AI systems can facilitate and support content moderation for humans[60] by flagging harmful content, blurring out areas that are particularly harmful, or engaging in 'visual question answering', i.e. humans' moderations can ask questions to the AI tool about the content without

53 Helberger, N. Karppinen, K., D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. In: Information, Communication & Society, 21:2, 191-207.
54 Content Personalisation Network. (2020). Retrieved on May 1, 2020, from https://www.projectcpn.eu
55 Papakyriakopoulos, O., Serrano, J.C.M., Hegelich, S. (2020). Political communication on social media: A tale of hyperactive users and bias in recommender systems. Online Social Networks and Media, 15. https://doi.org/10.1016/j.osnem.2019.100058
56 WFA and platforms make major progress to address harmful content. (2020). Retrieved on November 13, 2020, from https://wfanet.org/knowledge/item/2020/09/23/WFA-and-platforms-make-major-progress-to-address-harmful-content
57 Stolton, S. (2020). EU Rapid Alert System used amid coronavirus disinformation campaign. Retrieved on May 1, 2020, from https://www.euractiv.com/section/digital/news/eu-alert-triggered-after-coronavirus-disinformation-campaign/
58 Mozilla Insights. (2020). When Content Moderation Hurts. Retrieved on May 4, 2020, from https://foundation.mozilla.org/en/blog/when-content-moderation-hurts
59 Ofcom. (2020). Half of UK adults exposed to false claims about coronavirus. Retrieved on May 1, 2020, from https://www.ofcom.org.uk/about-ofcom/latest/features-and-news/half-of-uk-adults-exposed-to-false-claims-about-coronavirus
60 Ofcom. (2019). Use of AI in Online Content Moderation. Cambridge Consultants.

actually seeing it. The actual efficiency of these techniques also depends on the human response time to review the proposed content. Other AI-driven methods to tackle malicious online behaviour are to address the online audience directly in community management. Such AI 'nudging' techniques involve notifications or comments by chatbots that make the user aware that the post contains harmful content, or the technology can cause a short delay in the posting process which could encourage the user to rethink his or her message.[61] AI systems can also provide alternative, more positively expressed content suggestions which still resemble the original message. In both instances, the human agent, namely content moderator or user, takes the ultimate decision.

The complexity of online content challenges the *technical robustness* of AI systems in content moderation. First, AI systems face limitations due to the large variety of content formats, such as text, image, video, and audio which can also appear in a combination of different formats, such as in GIFs, memes, and emojis in combination with text. Advanced content types such as deepfakes and live video streams represent a considerable challenge for human and algorithmic content moderation.[62] Second, content moderation often requires evaluation beyond the content: it must take into account contextual understanding, e.g. societal, cultural, historical, and political aspects, and 'metadata', i.e. surrounding online information such as the number of followers and platform activities. Third, the variety of languages and nuances, e.g. sarcasm, represent challenges. These points create a challenge for both, algorithmic systems as well as human agents. However, users have higher expectations and less tolerance for mistakes in AI rather than human performance.[63] A poor algorithmic performance can have direct impact on the trust of humans in machines. To increase the *technical robustness*, training the AI systems requires large, suitable, and high-quality diverse datasets and constant updating, which is, however, challenged by complex contextual nuances. In particular, smaller newsrooms face difficulties in keeping up with big tech companies. Data and trained engineers for machine learning tend to be underrepresented and/or being insufficiently diverse, e.g. on gender, cultural background. In addition, training AI systems substantially affects *environmental well-being*. An AI training process is highly energy intensive and, hence, incurs considerable environmental costs. This leads to significant sustainability issues. [64] [65]

61 Statt, N. (2020). Twitter tests a warning message that tells users to rethink offensive replies. Retrieved on May 5, 2020, from https://www.theverge.com/2020/5/5/21248201/twitter-reply-warning-harmful-language-revise-tweet-moderation
62 Ofcom. (2019). Use of AI in Online Content Moderation. Cambridge Consultants.
63 Ofcom. (2019). Use of AI in Online Content Moderation. Cambridge Consultants.
64 Hao, K. (2019). Training a single AI model can emit as much carbon as five cars in their lifetimes. Deep learning has a terrible carbon footprint. Retrieved on May 31, 2020, from https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/
65 Matheson, R. (2020). Reducing the carbon footpring of artificial intelligence. MIT system cuts the energy required for training and running neutral networks. Retrieved on May 31, 2020, from http://news.mit.edu/2020/artificial-intelligence-ai-carbon-footprint-0423. bon footpring o

Transparency and *accountability* are two major principles to trace algorithmic decision-making and to counter potential abuse. First, the highly complex architecture of well performing AI moderating tools makes it difficult to analyse and reveal their decisions-making process.[66] Algorithmic standards that are not clearly defined and articulated can result in leaving 'negative' content online and/or removing 'appropriate' content. In this regard, it must be transparent who is to what extent accountable for algorithmic decisions and judgments and to whom it must be disclosed, e.g. general public, certain sectors, human agencies and/or oversight bodies. At the same time, transparency of algorithmic moderating tools towards users can increase the relation of trust between humans and machines. Second, a lack of transparency as well as the algorithmic system itself can be abused in political online campaigns during election periods, such as it often remains unclear who is paying for it, how much is being spent, and how audiences are segmented and targeted, e.g. through ads and chatbots.[67] An abuse of algorithmic filtering systems could further result in censorship which would violate democratic principles. In this regard Tracking Exposed[68] and Algorithms Exposed (ALEX)[69] introduced open-source software as algorithmic auditing methods to tackle the consequences of personalisation algorithms on social media and shopping platforms. Their goal is to empower both advanced users and low-skill users in the data extraction and enhance data literacy.

d. Theme 4: Automating communication

The fourth theme, **automating communication**, includes all forms of interaction and communicative actions and infrastructure enabled by AI. As such, chatbots, smart speakers, voice assistants, automated marketing communication belong to this theme. Everything from AI systems that simulate a proper conversation as well as encoding and decoding conversational messages and data from users falls under this theme. Referring to the 'Trustworthy AI' heptagon (Fig. 1), it is expected that the AI technology in this theme further diminishes aspect of human agents and is in favour of content generated by technological systems.

The most important key requirements for the automated communication theme are *human agency and oversight*, *diversity*, *non-discrimination and fairness* as well as *transparency*. First, transparency would require all AI-empowered communication channels to lay open or make auditable to specialists much of their data infrastructure and thus also how information and output is compiled. *Transparency* would also enable users to understand better how their conversation data is being used and evaluated. Especially in the field of automated marketing and communication, users can fall prey

66 EPRS. (2019). Understanding algorithmic decision-making: Opportunities and challenges.
67 Transparent Referendum Initiative. (2020). Retrieved on May 1, 2020, from http://tref.ie/
68 Tracking Exposed. (2020). Retrieved on May 1, 2020, https://tracking.exposed
69 Algorithms Exposed. (2020). Retrieved on May 4, 2020, https://algorithms.exposed

to misleading messages or biased information, which could be avoided if *transparency* in marketing practices would be made mandatory. More open and transparent automated communication technologies would ultimately give users greater reassurance while open infrastructures and datasets could enable research and generate public interest value. As such, open-source anonymous data may benefit for example the AI-driven development of translation services for low-resource languages. The *transparency* principle is therefore closely linked to maintaining *privacy and data governance*. At the same time, the trade-off by enhanced transparency could result in a backlash against data protection civil society groups advocating for better protection of aggregated datasets. In any case, ensuring multi-stakeholder governance of data and robust privacy measures is also relevant in relation to the data collection and purposes around voice and emotional AI, since it must be clear to users how the information is stored and used in a long-term perspective.

Referring to *diversity, non-discrimination and fairness* in automated communication AI systems, open unbiased datasets would not only favour users' communication experience but are also key to not distort a certain conversation or flow of information between humans and machines. Further, considering the EU's linguistic diversity, automated communication systems can already discriminate or disadvantage certain linguistic minorities. Finally, users should be able to choose whether they want to interact with a chatbot or with a human being, reflected in the principle *human agency and oversight*. This also links to the principle of *accountability* as far as imprecise or wrong information given by a consumer-oriented chatbot, e.g. for a bank, can cause harm or damage.[70] As such, the redressing of automated decisions by chatbots should be considered when discussing accountability in automated communication.

Also, *technical robustness* is relevant in that regard because the key principle would encourage more testing and development of automated communication systems prior to market them as a solution, by the creators to the potential customers. This would allow for a better user experience as well as more trust in automated AI-enabled communications.

The societal and environmental wellbeing principle also demands emphasis on the overall decision whether it is suitable, viable, sensible, considerate, and sustainable to adapt automated communication AI for a certain case. As such, beneficial cases include automated descriptions of visual content by using object recognition technology for the blind and vision-loss community.[71] Likewise, automated communication AI tools develop datasets and intelligent models that automatically translate online content

70 However, this also applies to wrong information from a human employee, and in both cases the bank is liable anyway.
71 Facebook automated alternative text. (2016). Retrieved on April 20, 2020, from https://www.facebook.com/accessibility/videos/1082033931840331/

for native speakers of low-resource languages[72], thereby making important content accessible to linguistically diverse communities. However, deploying already existing data risks replicating biases and errors from training datasets, e.g. stereotypes, gender and racial biases, especially for fully automated translation AI interfaces. The fact that employment opportunities for translators are significantly diminished by automated communication AI technologies threatens the *societal wellbeing* principle, according to which automated communication AI companies were required to mitigate the impact of their technologies on the traditional job sector.

Ultimately, automated communication should enhance human work which is achieved if the communication flows are still subject to human oversight. Delegating the decision to the AI system without human oversight should be avoided.

To improve the validity and emphasise the significance of the 7 Key Requirements in the four themes identified above, the figure below represents the view of the committee. The members assessed the significance of each requirement for Trustworthy AI within the context of the four themes of the MTS, i.e. automating data capture and processing, automating content generation, automating content mediation and automating communication. While the colours dark red and light red indicate a higher significance for the corresponding theme, orange, yellow and white reveal slightly lower significance. The view of the committee reveals that each of the 7 Key Requirements has a high significance throughout the MTS. This figure identifies *human agency and oversight*, *transparency*, and *accountability* as prevailing requirements throughout the MTS.

72 As such, an Irish research centre deploys automated translation and natural language processing AI for low-resource languages to preserve linguistic and cultural plurality in the EU: ADAPT center. Transforming Global Content. (2020). Retrieved on April 20, 2020, from https://www.adaptcentre.ie/research/transforming-global-content/

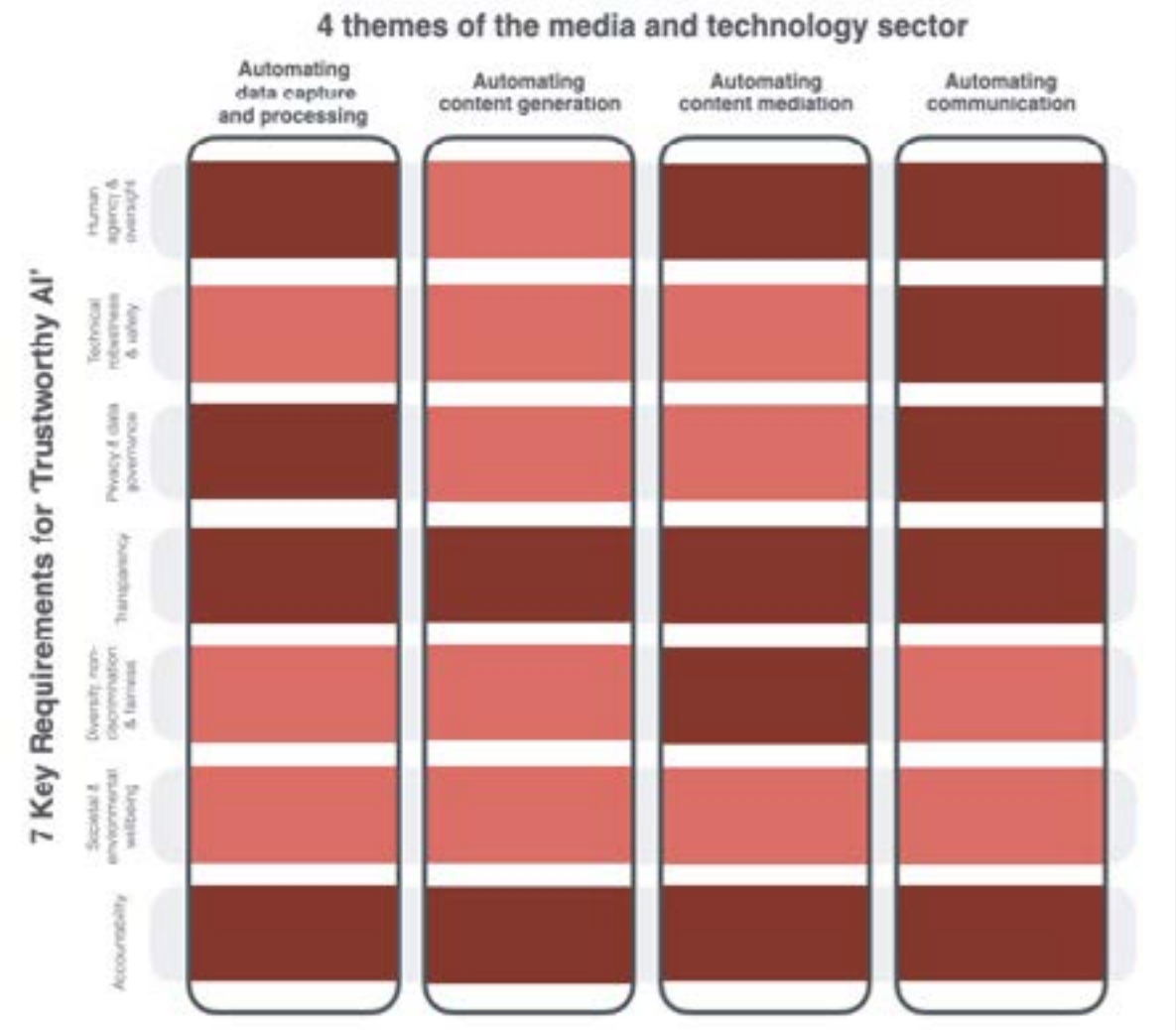


Figure 3: View of the committee on the significance of the 7 Key Requirements for 'Trustworthy AI' in relation to the four themes of the Media and Technology Sector

e. Possible tensions among the 7 Key Requirements for Trustworthy AI

Technical robustness and safety; Diversity, non-discrimination and fairness; Societal and environmental well-being

Possible **tensions** could arise between *technical robustness and safety*; *diversity, non-discrimination and fairness*; *societal and environmental well-being* because the large-scale implementation of AI tools such as holo-lenses for blind people as indicated in theme 1 (*automating data capture and processing*) still requires more extensive research and testing in order to be deployed on a large scale. Given the costs of development, it seems very hard to achieve non-discrimination in early adoption. In order to avoid longer term discrimination, care should be taken to ensure that products being developed

are trialled on diverse groups. In addition, currently, such advanced AI systems are not yet available for most of the blind population, especially for socioeconomically disadvantaged populations.

Technical robustness and safety; Human agency and oversight; Accountability; Transparency

Maximizing efficiency through *technical robustness* of AI systems in content generation can create **tensions** within the key requirements. Technical robust AI systems in data analysis can reduce *human agency and oversight*, *accountability*, and *transparency*. Furthermore, technical robust AI systems can become so efficient and well-advanced that they, ultimately, replace humans in tasks for which it is not necessary or desirable. This is at odds with *diversity, non-discrimination, and fairness and societal wellbeing*.

Transparency; Technical robustness and safety; Privacy and data governance; Societal and environmental wellbeing

Particularly in algorithmic content mediation, **tensions** can appear between *transparency*, *technical robustness*, *privacy and data governance*, and *environmental wellbeing*. A good moderating performance of AI systems might be based on a complex design of AI systems which, eventually, hampers explainability and transparency. Moreover, large datasets including a lot of user information are applied to increase the accuracy and efficiency of algorithmic systems. These data sets contain comprehensive information such as location, consumer preferences, political interests, education and workplace, relationship status, etc., which underlines once again the importance of privacy protection and data governance. Moreover, improving the accuracy of AI operations through well-trained system occurs at the expense of *environmental wellbeing*.

II. What must the Media and Technology sector do to be compliant with the 7 Key Requirements?

This section sets out guidelines for the implementation of AI in the Media and Technology sector. Specifically, it recommends how to adhere to the 7 Key Requirements, the *'Trustworthy AI' heptagon*, within the four identified MTS themes *automating data capture and processing, automating content generation, automating mediation, and automating communication*. Three clusters of recommendations are proposed: addressing data power and positive obligations (oriented mainly at people), empowerment by design and risk assessments (oriented mainly at infrastructure) and cooperative responsibility and stakeholder engagement (oriented mainly at stakeholders).

a) Addressing data power and positive obligations

Key requirements: *Privacy and data governance*; *Human agency and oversight*; *Transparency*

Aforementioned issues of consent are legitimate, particularly regarding the theme of *automating data capture and processing*. Do customers know when their personal data is being collected by AI-enabled systems? This relates to furthering 'data literacy' and 'data agency', which means stimulating awareness, building attitudes, enhancing capabilities and adjusting behaviour among users regarding (personal) data collection, processing and (re)use in the area of digital media and technologies.[73] However, at the same time, it should be avoided to put too much of the burden on the shoulders of relatively powerless citizens. It is first and foremost the task of data controllers to meaningfully explain what is happening with the data. Some users may never be fully digitally literate, yet data controllers also need to make clear to them what is going on. This requires more investigation into explaining well and meaningfully the data capturing, processing and (re)use. This could also mean a positive obligation for AI-driven business to conduct such research on an ongoing basis, as has been suggested in the past by WP29 in their guidelines on valid consent, which were recently updated by the European Data Protection Board.[74]

Positive data obligations also enable citizens to act with agency in the face of data power.[75] Automated data collection by AI systems happens in the background, particularly in remote biometric identification datasets and emotion detection AI. This raises, for example, the question if people should be able to decide if and how their emotions can be tracked, profiled, and re-used for specific purposes in order to avoid potentially harmful effects. For instance, Spotify's data analytics team conducts studies into musical preferences to profile users, not only to present them with better musical advice. One of Spotify's data analytics goals is to target advertising at users depending on the mood they are in, which is a play at manipulation using people's unconscious vulnerabilities.[76]

The meaningful, intentional and informed consent might erode in the presence of AI in the MTS. Users should, therefore, be informed when their volunteered, observed or inferred personal data is being used to train machine learning algorithms, and based on that decide whether to opt in, which could be described as **positive obligations**.

73 Pierson, J. (forthcoming) Media and Communication Studies, Privacy and Public Values: Future Challenges. In: González-Fuster, G., van Brakel, R. and De Hert, P. (eds.) Research Handbook on Privacy and Data Protection Law: Values, Norms and Global Politics, Cheltenham: Edward Elgar Publishing.
74 EDPB (2020). Guidelines 05/2020 on consent under Regulation 2016/679, adopted on May 4, 2020.
75 Kennedy, H., Poell, T. & van Dijck, J. (2015) Data and agency. In: Big Data & Society, July-December, 1-7.
76 See e.g. https://mitpress.mit.edu/books/spotify-teardown

Therefore, **the Committee recommends ensuring clear and strong consent (opt-in) and transparency obligations** for algorithmic training and testing with user data in MTS. This can be operationalised by for example setting-up algorithmic registries, as done by the cities of Amsterdam and Helsinki.[77] On top of providing understandable and easily accessible information on *automating data capture and processing* to users, the latter must also be able to contact a human to provide further information about the aforementioned aspects and users must be guaranteed satisfactory and effective remedies if they have been negatively affected by decisions of AI systems.[78] **The Committee, therefore, recommends responsive redress mechanisms.**

Disclosure of personal data should be a human-consented transaction, not one enticed or (unconsciously) demanded by technology. Data minimisation by design as required by the GDPR should be clearly implemented and enforced in the MTS. Companies should be obliged to undergo regular data reviews to ensure they are not 'casting their nets' farther than necessary. Lastly, data anonymisation or at least pseudonymisation by design should become a key principle. More research investments by the MTS sector are needed in this field. Pseudonymising data is not only favourable for users, but further mitigates risks arising from data breaches, systemic surveillance and cybercrime.

Explainability is a complex, nuanced problem, considering the variety of European citizens. Research and funding for increasing AI transparency and explainability should be pursued and prioritized. This should be combined with (co-)regulatory efforts for establishing more transparency from digital platforms vis-à-vis independent regulators, on matters like internal processes for handling harmful and illegal content through algorithms and AI. In that way we can better address and regulate the behaviour of platform-specific architectural amplifiers of contentious content, e.g. in recommendation engines, search engine features (such as autocomplete), features like 'trending', and other mechanisms that predict what we want to see next. This approach fits in with suggestions being made on ex ante principles-based co-regulatory approaches for addressing online harms as a key operational objective of digital platforms, in a way which is reflective of their reach, their technical architecture, their resources, and the risk such content is likely to pose.[79] Hence, **the Committee recommends strengthening research, process-based (co-)regulation and oversight on AI transparency and explainability**, especially with regards to architectural elements for algorithmic amplification.

77 Moltzau, A. (2020). Algorithm Registries in Amsterdam and Helsinki. Retrieved on November 13, 2020, from https://alexmoltzau.medium.com/algorithm-registries-in-amsterdam-and-helsinki-c1364b70ca6

78 Fanni, R., Steinkogler, V. E., Zampedri, G., & Pierson, J. (2020, September). Active Human Agency in Artificial Intelligence Mediation. In Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good (pp. 84-89).

79 Vermeulen, M. (2019). Online Content: To Regulate or not to Regulate-Is that the Question?. Vermeulen, Mathias, Online content: to regulate or not-is that the question.

Anticipatory data management policy should be a future priority in EU legislation. Privacy is a moving target, and new categories of personal data will be utilized, collected and created. Therefore, it is imperative that GDPR and the ePrivacy directive update consider emerging sensitive AI-related personal identifiers, whether emotional data or even predicted behaviour AI systems foresee an individual taking.

Individual consent decisions will not prevent all types of societal harms stemming from abusive uses of automated personal data processing. While individuals may consent to the use of information about e.g. their emotions, political affiliation, health or sexual orientation, this may have large-scale effects beyond a single citizen, for which individual choices cannot bear responsibility. Political microtargeting offers an example: individual users may consent to the use of data about their political preferences and emotional states on a platform, but in aggregated form, data on attitudes and emotions linked to political preferences may be used to automatically manipulate voting behaviour of other citizens with potentially major societal effects, as the Cambridge Analytica scandal has illustrated.[80] Prevention of such malignant applications of automated data processing cannot rest on an individual's shoulders and should be addressed with regulation based on an interdisciplinary, multi-stakeholder engagement to uphold public values.

The Committee recommends multi-stakeholder processes for investigating how predictive analytics, sentiment analysis and emotional AI threaten the integrity and autonomy of digital media users, especially in online behavioural advertising and synthetic content production. This approach is in line with the remit of Art. 22 GDPR ('Automated individual decision-making, including profiling').

b) Empowerment by design and risk assessments

Key requirements: *Human agency and oversight*; *Transparency*; *Diversity, non-discrimination and fairness*; *Societal and environmental well-being*; *Technical robustness and safety*

AI technologies being used for activities like profiling, content personalisation and targeted advertising can pose threats to *human agency*, to *transparency*, to *diversity, non-discrimination and fairness*, to *societal well-being*, and to *technical robustness and safety*.

Therefore, it is important that comprehensive solutions are being investigated and developed to address these threats. This fits in with the idea of 'empowerment by design', i.e. building infrastructures and systems in such a way that (organised) citizens have agency to safeguard and strengthen their fundamental rights and the public interest.[81]

80 The Guardian, The Cambridge Analytica Files. Retrieved on November 13, 2020, from https://www.theguardian.com/news/series/cambridge-analytica-files

81 Pierson, J. and Milan, S. (2017) Empowerment by design: Configuring the agency of citizens and activists in digital infrastructure. Presentation at Communication Policy & Technology section for IAMCR Conference 'Transforming Culture, Politics & Communication: New media, new territories, new discourses', 17 July 2017, Cartagena, Colombia.

The targeted advertising industry in MTS is a complex and multi-sided market with a multitude of actors, many of whom intermediaries, such as networks of third parties with tracking technology, intermediary data brokers, and exchanges all competing in the market of RTB and automated auctions.[82] Sensitive information about individuals can be inferred and used, e.g. ethnicity, gender, sexual orientation, religious beliefs, for online behavioural advertising and affinity profiling, i.e. grouping people according to their assumed interests rather than their personal traits. Several scholars and digital rights organisations have made suggestions for empowering consumers in case of illegal or unethical automated capturing and processing of their personal data. Hence **the Committee recommends investigating comprehensive solutions for addressing legal and ethical risks of automated decision-making and profiling, like the 'right to reasonable inferences'**.

Besides issues of profiling in digital marketing, AI is also used in emotion detection and sentiment analysis in MTS. This can have positive uses, but it also bears risks to manipulating human behaviour. These systems could powerfully 'nudge' people into taking certain behavioural actions; used to infer belief and attitude; and incentivise use or concealment of certain emotional expressions. Emotion detection could likewise exacerbate existing biases specifically for vulnerable groups of the society. A set of actions could help to mitigate the risks posed by emotion detection AI. First, users should have to opt-in if any of their data is being used to detect emotions. The consent by users should be mandatory for MTS business, as required by EU data protection law. However, consenting to the data collection does not suffice, as the issue lies with how the results of data analysis are applied, e.g. avoiding that citizens are manipulated at scale. The (dynamic) consent should be reviewed and renewed on a recurring basis with full disclosure over the purpose and scope of the emotion AI implementation areas, and only for sound reasons such as health or safety. Those developing sentiment analysis and emotion detection AI need to be urged to full transparency and public discussion with relevant experts such as sociologists, psychologists, anthropologists, media scholars and psychiatrists. Overall, **the Committee recommends designing an EU-wide, dynamic, and mandatory high-risk assessment scheme** for AI systems detecting sentiments from their users, leading to empowerment by design for citizens and society.

82 Binns, R., Zhao, J., Kleek, M. V., & Shadbolt, N. (2018). Measuring Third-party Tracker Power Across Web and Mobile. ACM Trans. Internet Technol., 18(4), 52:1–52:22. https:// doi.org/10.1145/3176246

More largely, high-risks assessment schemes also need to consider the value of the AI-enabled system(s) against the risks. The latter also refers to minimising unintentional and unexpected harm, and preventing unacceptable harm, which is related to the principle of technical robustness and safety. Simply put, the value of the service enabled/provided must outweigh the risk of the data collected. Thus, a theoretical continuum exists where risks associated with disclosure of personal data and reward or value of received product or service are balanced cognitively.[83] [84] This applies in scenarios where humans interact with AI systems, such as in the first theme. The AI HLEG Assessment List for Trustworthy Artificial Intelligence (ALTAI) already provides a tool to self-assess compliance of specific AI use cases with the 7 Key Requirements for Trustworthy AI. The Committee recommends that the EU-wide, dynamic, and mandatory high-risk assessment scheme should be coherent with the ALTAI, specifically focussing on the potential risks and societal impacts arising in the MTS.[85]

c) Cooperative responsibility and stakeholder engagement

Key requirements: *Accountability*; *Societal and environmental well-being*; *Diversity, non-discrimination, and fairness*; *Transparency*

Many concerns that arise in this sector can only be tackled by means and resources beyond the sector. For instance, social media has enabled targeted harassment of private individuals, which may evade current attempts to regulate, and savvy abusers can readily avoid penalty. *Accountability* issues can arise if companies fail to catch up with technology, if the technology or service provided is ineffective, or if services available only to people with plenty of resources. Yet, effective legal remedies against abusive individuals could be one way of helping to prevent blanket social media policies which may have more draconian effects on freedom of expression. Targeted online harassment of individuals needs to be taken more seriously especially considering the EU fundamental human rights framework and legal obligations. These policies should consider the context since it is vital to communication and hence, a policy that works in one context in social media could be disastrous in another.

ICT blurred borders between media production, consumption and literacy. The most effective way to secure *societal and environmental well-being* should be a shared responsibility between civil society (users), industry (platforms) and governments (education remit). This type of 'cooperative responsibility' requires that digital media platforms, policy makers, users and possible other actors develop a division of labour

83 Robinson, C. (2017). Disclosure of personal data in ecommerce: A cross-national comparison of Estonia and the United States. Telematics and Informatics, 34(2), 569-582.
84 Petronio, S. (2002). Boundaries of privacy: Dialectics of disclosure. Suny Press.
85 European Commission. (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. Retrieved on July 17, 2020, from https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

on how to manage their responsibility for their role regarding public values.[86] The EU preliminary principle demands that the MTS can only be 'compliant' in presence of an oversight body including a transparent system of compliance, an appeal (redress) and a complaints procedure. Any such system would also have to acknowledge and interface somehow with legacy governance structures in the MTS. Given the legal obligations in the EU, **the Committee recommends setting up an advisory body with all relevant stakeholders involved for feedback and evidence on EU technology policy**.

Providing an outlook, the New European Media Initiative (NEM)[87] is a key European technology platform organisation for the MTS, that – since Framework Programme 7 – is intensely involved in the EU research, thereby driving the future of digital experience. In their "Vision Paper 2030 – Towards a future media ecosystem", NEM aims to unite the MTS with European core values, drivers and goals. Acting ethical, transparent and accountable, being human-centric and sustainable, and encouraging an empowered and critical society are the main ambitions.[88] **In line with the 7 Key Requirements for AI in the MTS, the Committee recommends fostering exchanges and best practices with other institutions, network organisations and multi-stakeholder initiatives**, for example, NEM, Forum on Information & Democracy,[89] Re-Imagine Europe,[90] and the Council of Europe.

Furthermore, impacts on human creativity and societal wellbeing in the media and creative industry could be serious, e.g. in case of music creation by AI. Remedies could, for example, include tax channelled to live music venues/music schools, regulators to remove barriers to live music performance, encouragement of music tuition at all levels of schooling, and open provision of software to educational establishments. This also includes broader support for public service media and creative industry to safeguard creativity and wellbeing. Therefore, **the Committee recommends allocating funding to the most severely impacted creative media industries in the EU**, especially on cultural and public service/information grounds.

The MTS and online intermediaries in particular should be encouraged more to set up an appropriate architecture for empowering users. More standardised methodologies and deliberation fora to facilitate ongoing exchange with the specific user community should be put in place. Also, media production cycles such as designing websites (access, monitoring and dissemination) should involve multiple stakeholders. Likewise, the same stakeholders should be taught the essentials of *diversity*, *non-*

86 Helberger, N., Pierson, J. and Poell, T. (2018) Governing online platforms: from contested to cooperative responsibility. In: The Information Society, 34 (1), 1-14.
87 https://nem-initiative.org
88 Adzic, J., D'Andria, F., Behrmann, M. Boi, S., Castillo, P., Clarke, J., Danet, P-Y., Delaere, S., Fernandez, S. Hrasnica, H. Lippold, S., Matton, M., Menéndez, J.M., De Rosa, S. (2020) NEM Vision 2030: Towards a future media ecosystem, NEM – New European Media, April 2020, 18.
89 https://informationdemocracy.org
90 https://reimagine-europa.eu

discrimination, fairness and human rights, as the ISFE-Council of Europe guidelines to online game developers did.[91] **The Committee recommends incentivizing and developing educational trajectories, guidelines, training, materials and tools** for professional and technical staff (e.g. via online courses or curriculum changes in higher education) to better understand and engage with EU fundamental human rights and the principle of trustworthy human-centered AI.

Example: The PEGI Case

To present the recommendations in applied context, the Pan European Game Information (PEGI) System demonstrates how a voluntary regulatory system can work in practice. The system recommends content and age policies for video games. It is pan-European, interacts with other regional systems in Asia and North America, and sits on top of national governance systems. PEGI is advised by national councils and an expert advisory board made up of representatives (e.g. academics, parent bodies, film rating bodies) from around Europe. These all meet with PEGI staff face to face once a year and online in between the annual meetings. The committee member names are published online, which provides transparency. PEGI and its North American and Asian equivalents are working together to develop an International Age Rating Coalition (IACR).

PEGI is a system that results in information notices on the back of physical boxed media products and now also in the online app and other stores. Publishers fill out a questionnaire and send it to PEGI before a game is released. PEGI can refuse to give a rating to a game, ask for clarifications and it can increase or decrease a rating on appeal. It also takes complaints directly from the general public.

The system works reasonably well in terms of a high level of accountability, but it also has weaknesses. Some online platforms do not participate. How games are rated and on what grounds can be opaque to those outside of the organisation. Further, the system does not have legislative backing and thus cannot take punitive actions against game companies like the game rating systems for example in Germany and the UK do. Thus, while under national legislation it is illegal to sell an over 18 game to a minor in the UK, this is a matter of national legislation. The system is highly focused on protecting children but less on negative impacts or procedures for adults or other vulnerable populations. Further, it is unclear what impact the system has in practice in terms of purchasing behaviour and game playing. PEGI is a co-regulatory system, with a focus on 'educating' consumers but especially protecting children. For a critical discussion see Felini (2015).

91 DG of Human Rights and Legal Affairs. (2008). Human rights guidelines for online game providers. Developed by the Council of Europe in co-operation with the Interactive Software Federation in Europe. Retrieved on June 1, 2020, from https://rm.coe.int/16805a39d3.

Any system that might emerge may want to consider the rather stronger role and stance taken in some countries in relation to the 'traditional media' industries including for example the Press Councils and Press Ombudsman in Ireland which operates to oversee both print and online only news media[92] and the communications regulation bodies like Ofcom in the UK which oversee telecoms and broadcast media.[93] Any governance system might also need to work with established worker unions like the National Union of Journalists, both in terms of training and educating journalists, and in terms of whistleblowing and worker rights. In sum, **the Committee recommends strengthening workers' rights and public interest values in the media as new AI systems evolve and emerge**.

Public information campaigns and initiatives about the functioning and possible risks of new AI initiatives should be promoted. As such, the Media Literacy Initiative[94] involves public, commercial and not for profit/community organisations to counter mis- and disinformation around Covid-19 and is running across online and traditional media channels.[95] Similar information and public communication initiatives are taken at European level including of course Safer Internet Day.[96] **The Committee recommends extending existing publicly supported media, data and AI literacy programmes to include information and public awareness of AI applications, services and impacts**.

5.
Conclusion

Artificial intelligence systems have a substantial impact on various areas of the European media and technology sector (MTS). This report identified four themes of AI applications in the MTS: *automating data capture and processing, automating content generation, automating content mediation, and automating communication*. This report analysed the core opportunities and risks of AI applications within these proposed themes. The 7 Key Requirements for Trustworthy AI developed by the European Commission High-Level Expert Group on AI were at the centre of discussion. The report addresses its recommendations to all stakeholders involved in the development, deployment, use, and governance of AI systems in the MTS.

92 Press Council of Ireland. Office of the Press Ombudsman (2020). Retrieved on June 1, 2020, from https://www.presscouncil.ie/.
93 Ofcom. (2020). TV, radio and on-demand. Retrieved on June 1, 2020, from https://www.ofcom.org.uk/tv-radio-and-on-demand.
94 Be smart media. An Initiative of Media Literacy Ireland. (2020). Members. Retrieved on June 1, 2020, from https://www.bemediasmart.ie/members.
95 Be smart media. An Initiative of Media Literacy Ireland. (2020). About. Retrieved on June 1, 2020, from https://www.bemediasmart.ie/about.
96 Be smart media. An Initiative of Media Literacy Ireland. (2020). Members. Retrieved on June 1, 2020, from https://www.bemediasmart.ie/members.

Recommendation cluster 1: Addressing data power and positive obligations

- Ensuring clear and strong consent (opt-in) and transparency obligations for algorithmic training and testing with user data in MTS.
- Establishing responsive redress mechanisms, so that users can contact humans to provide understandable and easily accessible information on automating data capture and processing, and have satisfactory and effective remedies when negatively affected by AI decisions.
- Strengthening research, process-based (co-)regulation and oversight on AI transparency and explainability, especially with regards architectural elements for algorithmic amplification.
- Ensuring a multi-stakeholder process for investigating how predictive analytics, sentiment analysis and emotional AI threaten the integrity and autonomy of digital media users, especially in online behavioural advertising and synthetic content production.

Recommendation cluster 2: Empowerment by design and risk assessments

- Investigating comprehensive solutions for addressing legal and ethical risks of automated decision-making and profiling, like the "right to reasonable inferences".
- Designing an EU-wide, dynamic, and mandatory high-risk assessment scheme for AI systems detecting sentiments from their users, leading to empowerment by design for citizens and society.
- The EU-wide, dynamic, and mandatory high-risk assessment scheme should be coherent with the ALTAI, specifically focussing on the potential risks and societal impacts arising in the MTS.

Recommendation cluster 3: Cooperative responsibility and stakeholder engagement

- Setting up an advisory body with all relevant stakeholders involved for feedback and evidence on EU technology policy.
- Fostering exchanges and best practices with other institutions, network organisations and multi-stakeholder initiatives.
- Allocating funding to the most severely impacted creative media industries in the EU, especially on cultural and public service/information grounds.
- Incentivizing and developing educational trajectories, guidelines, training, materials and tools for professional and technical staff to better understand and engage with EU fundamental human rights and the principle of trustworthy human-centered AI.
- Facilitating and strengthening workers' rights and public interest values in the media as new AI systems evolve and emerge.

- Extending existing publicly supported media, data and AI literacy programmes to include information and public awareness of AI applications, services and impacts.

The report concludes by emphasising the involvement of public, private, scientific and civil society stakeholders in order to achieve a holistic AI governance framework across the EU.

This report and especially the proposed recommendations aim to tackle concerns that arise due to the proliferation of AI systems in the MTS, thereby ensuring an ethical and sustainable AI implementation throughout this sector. As such, public, private and civil society organisations representing the media and technology sector in Europe, as well as other institutions in Europe are encouraged to consult this report and actively implement the proposed recommendations.

Acknowledgements
We thank Valerie Eveline Steinkogler and Rosanna Fanni for their assistance with the development of the project and for contributing to the draft report, as well as Giulia Zampedri for the support in the finalisation of the report. In addition, we are also grateful to Ana Pop Stefanija and Ine van Zeeland for their revisions and input, in their capacity as PhD researchers for respectively the FWO Research Project DELICIOS (Delegation of Decision-Making to Autonomous Agents in Socio-Technical Systems) (https://coast.uni.lu/delicios) and the VUB Research Chair 'Data Protection on the Ground' (www.dataprotectionontheground.be).

ATOMIUM
EUROPEAN INSTITUTE
FOR SCIENCE, MEDIA AND DEMOCRACY

From left to right: Valéry Giscard d'Estaing, Jean-Claude Juncker and Michelangelo Baracchi Bonvicini.

Atomium-European Institute for Science, Media and Democracy (EISMD), convenes leading European universities, media, businesses, governments and policymakers to increase the exchange of information and interdisciplinary collaboration, to develop innovative collaborative initiatives and to encourage frontier thinking about science, media and democracy.

Atomium-EISMD was launched publicly by the former President of France Valéry Giscard d'Estaing, Michelangelo Baracchi Bonvicini and by the leaders of the institutions engaged during the first conference on the 27 November 2009 at the European Parliament in Brussels.

With the contribution of:

INTESA SANPAOLO ELSEVIER facebook FUJITSU Microsoft ZURICH