The Role of Data in AI

Report for the Data Governance Working Group of the Global Partnership of AI

Report prepared by:

Digital Curation Centre
Trilateral Research
School of Informatics, The University of Edinburgh

November 2020

Co-Chairs' Foreword - The Role of Data in AI

The Global Partnership on AI (GPAI) was founded with a mission to "support and guide the responsible adoption of AI that is grounded in human rights, inclusion, diversity, innovation, economic growth and societal benefit, while seeking to address the UN Sustainable Development Goals (UN SDGs)" and "facilitate international project-oriented collaboration in a multistakeholder manner with the scientific community, industry, civil society, international organizations, and countries".

When we were invited to become the Co-Chairs of the Data Governance Working Group, we were delighted to be asked to support this mission. We welcomed its focus on practical impact, and recognised that good data governance - collected, used and shared in responsible and trustworthy ways - will be foundational to this ambition and many of GPAI's future projects.

That is why we were excited to commission the Digital Curation Centre and Edinburgh University's School of Informatics alongside Trilateral Research as a consortia to help identify concrete areas for international collaboration, including areas where more data would be useful – such as specific, open, datasets that could be worthy of further support – and where harms arise due to the collection of or access to data. The team acted independently from the Working Group, but consulted its members, as well as its Steering Committee, in the course of its mandate.

The report provides a deeper investigation into many of the areas discussed in the Working Group's Framework and has been produced in parallel in preparation for the Summit. It reflects the technical and legal expertise of the consortia, and its recommendations - highlighting specific initiatives that could advance GPAI's mission - will help inform the next phase of the Working Group's work as we identify projects and programmes of work that align with GPAI's mission, and could be funded by GPAI's members and in partnership with others.

The Working Group has a mandate that aligns closely with GPAI's overall mission: to "collate evidence, shape research, undertake applied AI projects and provide expertise on data governance, to promote data for AI being collected, used, shared, archived and deleted in ways that are consistent with human rights, inclusion, diversity, innovation, economic growth, and societal benefit, while seeking to address the UN Sustainable Development Goals." We thank the Digital Curation Centre, School of Informatics and Trilateral Research for their dedicated work and contribution to the practical realisation of that vision.

Dr. Jeni Tennison
Vice-President and Chief Strategy Adviser
Open Data Institute

Dr. Maja Bogataj Jančič
Founder and Head
Intellectual Property Institute

Co-Chairs of the Data Governance Working Group

Executive Summary

This is the final report of the project *The Role of Data in AI,* which was commissioned by the Data Governance Working Group (WG) of the Global Partnership of AI (GPAI). The overarching aim of the project was to highlight and describe the role of data in AI development processes and identify key challenges related to data data quality, accessibility and availability. We also describe the impact these challenges have on AI development, at societal and individual levels.

The Role of Data in AI project ran between 17th September - 7th December 2020 and was led by the Digital Curation Centre, with project partners Trilateral Research and School of Informatics, The University of Edinburgh. The report is based on a review of literature and consultation with expert members of GPAI and the Data Governance WG through a series of three workshops and weekly meetings.

The first three sections of the report describe the role of data in AI development as well as key types of data that are used and their characteristics. It highlights the importance of having vast amounts of good quality data for AI development for best results and how data limitations can lead to poor results, which can have negative impacts on society and individual rights. Section 5 goes into more depth and examines data-related issues emerging from the collection, process and use of data in AI and offers a wide mapping of important issues to inform the further developments of AI creation. It provides a brief examination of the impact of access to datasets and use of different types of data for the creation of AI.

Section 6 outlines the impact of the law on access to and availability of data, indicating the complexity, challenges and risks of global privacy or IP legal regimes. It concludes that establishing concrete rules to govern data sharing among public and private actors can benefit the development of AI and increase the flow and availability of data. These actions will however take concerted efforts from policy makers and legislators to be realised.

Section 7 draws on three case studies to examine context-specific data challenges in three fields: Development of Human Language Technologies, AI for Pandemic Response and the use of AI in the Criminal Justice System. Although these cases differ, and subsequently the challenges data poses, there are important issues that have been identified across the three examples outlined. The lack of access to data, data availability and quality can have far reaching harmful impacts, ranging from lack of available services in communities' native languages to racial bias in sentencing. AI has great potential to assist in response to pandemics, due to its ability to analyse vast amounts of data in short time; however, as the case study shows, lack of quality and harmonised data has hampered the ability to utilise AI to its full potential in the current response to COVID-19.

Many of the challenges identified in the report can be mitigated or overcome through data governance and the report concludes with recommendations to the Data Governance WG on priority areas to work on their ongoing work within GPAI, with a focus on data quality, access and availability, to ensure that the benefits of AI can be more evenly realised across the globe.

This report was prepared by staff of the Digital Curation Centre, Trilateral Research and The School of Informatics, University of Edinburgh and was submitted to the Data Governance WG on 26th November 2020.

Key contact person is:

Dr Thordis Sveinsdottir, Senior Research Data Specialist
Digital Curation Centre
thordis.sveinsdottir@ed.ac.uk

This report was commissioned by experts of the Global Partnership on Artificial Intelligence's Working Group on Data Governance. The report does not necessarily reflect the views of the experts' organizations, GPAI, the OECD or their respective members.

Table of Contents

Co-Chairs' Foreword - The Role of Data in AI 2

Executive Summary 3

1. Introduction 7

2. AI Development and the Role of Data at Each Step 9

2.1 Data collection and creation 9

2.2 Data organisation/refinement 10

2.3 Learning from data 11

2.4 Evaluation 12

2.5 Retention/preservation of data sets 12

2.6 Deletion 13

3. Data Types Used in AI Development 14

3.1 Data states 14

3.2 Data types 15

4. Data Characteristics that Influence the Process or Outcome of AI Development 19

4.1 Data quality and data governance 19

4.3 Representativeness 20

4.4 Accuracy 21

4.5 Completeness 22

4.6 Accessibility 23

4.7 Coverage 24

5. Socio-Ethical, Economic and Environmental Impact of Data in AI 26

5.1 Socio-ethical impact 26

Autonomy, beneficence, non-maleficence and justice 26

Inclusion and exclusion 28

Equality and non-discrimination 29

Racial discrimination 29

Gender discrimination 30

5.3 Economic impact 30

5.4 Environmental impact 32

6. Law and Transparency as Modifiers to Impact of Data in AI 35

6.1 Impact of law on data availability and data access 35

6.2 Transparency for data and AI 38

7. Availability of and accessibility to data for AI development: data quality and challenges in three fields 42

7.1 Development of AI in the Pandemic Response 43

Health data for developing rapid diagnostic AI in a pandemic 43

Chemical and drug data for AI assisted drug discovery/drug repurposing 45
7.2 Human Language Technologies for under-resourced languages 47
7.3 Data for developing AI applications in the Criminal Justice System 49
7.4 Data management for supporting AI development across different fields 50
Recommendations on data management within AI projects 51
7.5 Unavailable data, legal and commercial challenges 52
8. Recommendations to the Data Governance WG for work on data governance for AI 54
9. Concluding Remarks 57
Resources 58
Annex A- Initiatives and projects working on challenges relating to data governance, availability and accessibility. 70

1. Introduction

'The availability of data is essential for training artificial intelligence systems, with products and services rapidly moving from pattern recognition and insight generation to more sophisticated forecasting techniques and, thus, better decisions. [...] Moreover, making more data available and improving the way in which data is used is essential for tackling societal, climate and environment-related challenges, contributing to healthier, more prosperous and more sustainable societies.' (European Commission, 2020:2-3)

This report on *The Role of Data in AI* is written for the Data Governance Working Group of the Global Partnership on AI (GPAI). The WG has the mandate to 'collate evidence, shape research, undertake applied AI projects and provide expertise on data governance, to promote data for [Artificial Intelligence (AI)] being collected, used, shared, archived and deleted in ways that are consistent with human rights, inclusion, diversity, innovation, economic growth, and societal benefit, while seeking to address the UN Sustainable Development Goals.'[1]

As part of the group's work, this report was commissioned to highlight the importance of data for AI, and how unavailability and lack of accessibility to good data sources negatively impacts on the development of AI applications. AI is currently being used to accelerate progress in vital fields such as health, agriculture, finance and transport. For AI to succeed in furthering development evenly across areas and regions, good data sources are vital. As of yet, data gaps still exist, and where data is available, there may be challenges with access and/or data quality can be poor.

The report is based on findings from a literature review and results from three expert workshops held with the Data Governance WG in October and November 2020. Extensive scoping and review of literature from academic, grey and government sources was carried out to illustrate the role of data in AI and highlight the current key challenges with regard to data governance, access and availability. The review was also focused on finding best practices and work currently being undertaken internationally to overcome these challenges.

The report presents an analysis of the challenges and provides recommendations and examples of best practices to assist the Data Governance WG, as part of GPAI, in their ongoing mission to support good data governance for AI projects and systems.

This report is divided into 7 main sections:

Section 2: Outlines key steps in the use of data from AI development from data collection/creation to preservation/deletion.

Section 3: Describes the main types of data that are used for AI development and how the availability of data types influences AI development. We also look at how the specific requirements of AI can play a role in the demand for certain types of data.

Section 4: Describes the important characteristics of data that influence the process or outcome of AI development. This section explores the concept of data quality and illustrates its importance for the development of relevant and unbiased AI technologies.

[1] See Foreword to this report.

Section 5: Examines the impact of unequal access to datasets and use of different types of data for the creation of AI. Benefits as well as potential harms on socio-ethical, economic, environmental and legal levels are identified and discussed.

Section 6: Discusses the impact of the law on the access to and availability of data in the creation, development and employment of AI indicating the complexity, challenges and risks of global privacy or IP legal regimes.

Section 7: Carries on the discussion of characteristics with a focus on describing data quality and data challenges in three case studies of AI development: Development of Human Language Technologies for Under-Resourced Languages, Development of AI for Pandemic Response and Use of AI in the Criminal Justice System. This section further illustrates and provides recommendations on how good data management can assist with mitigating these challenges.

Section 8: Draws on work in the previous sections to present a set of recommendations to the Data Governance Working Group on how to further data governance for AI data to drive standards around data quality, discoverability, availability and accessibility.

2. AI Development and the Role of Data at Each Step

This section will outline key steps in the use of data for AI development, from the perspective of a project creating a new AI product, from data collection to preservation or deletion. It will, at each step, offer insights into data related challenges and offer examples to further describe the processes and any barriers faced by AI developers.

Building an AI system typically involves sourcing large amounts of data and creating data sets for training, testing and evaluation, and then deployment. This process is iterative in the sense that it may require several rounds of training, testing and evaluation until the desired outcome is achieved and data plays an important role at each step. This report will not go into the details of the inner workings of an AI system but follow the journey of the data through the system development cycle.

2.1 Data collection and creation

The first step in building an AI system is considering the problem it has to solve. Data availability will have a major impact on how the system is assembled and what AI techniques will be used. The quantity and quality of data available will have an impact on the quality of the final product. In this sense, one can argue that data availability (whether data exists) and accessibility (whether data is accessible) are the main driver behind development of products that use AI technologies.

AI is used in many products that have a positive impact on the quality of life, for example, speech technologies, tools for trading and investment, development of medicine, law enforcement, environmental forecasting, products such as self-driving cars, etc. Unfortunately, lack of access to data in certain domains can potentially result in the entrenchment of inequalities or under-representation of particular groups or communities. For example, lack of data in most of the world's languages means that most speech technologies are overwhelmingly available in European languages (Besacier, 2014). Many communities lack access to important developments in health, education, public services and finance, which serve to further amplify existing inequalities and create new ones.

Sourcing data can be very difficult, and many projects must create a data corpus from scratch, which can be costly. Crystal (2000) estimates that producing data for speech technologies in a language can cost $80,000.[2] Organisations that undertake such investment on data creation may be reluctant to share it for free and this impacts on overall data accessibility for AI development. Academic institutions are typically more open to sharing the data they collect as increasingly this is a requirement associated with research supported by public funds. However, they tend to operate on inside knowledge about the latest research in their field, and data may not be discoverable by external developers as data sources are often not centralised and findable. In some fields, cost savings can be a driver for building data sets. For example, AI tools may be more efficient than humans at diagnostics based on medical imaging so investing in a data set of images may be very cost-effective in the long term.

Given that recent innovation in AI has been largely data-driven, this has also resulted in a data economy. It is important that data is available to both academic researchers and businesses, as both sectors do valuable work in the field. However, it is estimated that "the big 5" (Apple, Amazon, Facebook, Google, Microsoft) have a combined value of 6 trillion USD (Scelata et al, 2019). There are concerns that oligopolies, or even monopolies, are emerging in certain areas

[2] We note that this amount will be depending on the availability of data, which varies considerably between languages, as well as cost of digitisation, community and government support.

(Moore, 2019). Users (data contributors) are typically asked for permission, by signing privacy agreements, nonetheless criticisms around transparency in this area are longstanding as it is not always clear to users what the data will be used for. Data protection and privacy legislation and regulation in countries across the globe has been passed to address these issues and we will revisit this in section 5.

Data misuse must be prevented to maintain public trust in AI systems and the decision-making it informs. In addition to legal protection, the public should also be provided with digital education, including information on data ownership and about AI as a science, in general. Public involvement is valuable in building data corpora. For example, people have donated voice recordings to public data corpuses, which are extremely valuable in building speech technologies in a variety of languages. Another valuable example of public data input is through the use of track and trace applications to enable authorities to build intelligence on the spread of the COVID-19 pandemic.

2.2 Data organisation/refinement

Once it has been established what data is available for a project, the next stage involves refining the final product by assessing and deciding whether further data is needed.

For example, to create a speech synthesis system, at a very minimum the developers require a pronunciation lexicon, text script and recorded sound files, which can be used to create a voice. They must decide if the data measures up to the goal. If the voice is intended for commercial sale, they may need to employ a voice talent to obtain better quality recordings. The good news is that open-source data are being built with crowd participation, such as Mozilla Common Voice (Wiggers, 2019). For certain purposes, it may be useful to collect publicly available text, audio-video files by automatically scraping them from the Internet, typically from various news media as well as social media.

With respect to data, quantity does not necessarily equal quality. Of most importance is that existing sound data is of good quality, is representative and has good coverage, e.g. the script covers as much as possible of the lexicon and sound files cover as many sound sequences as possible (see more on data quality in Section 4).

The next step is cleaning the data and getting it ready for training. This can be a time-consuming process and it involves removing data that is likely to skew results, retaining enough noise in the data to avoid overfitting. It is important to get to know the data well and how to best organise it to enable learning. Not understanding how the data was put together and how it works with the processing models, will hinder the effectiveness of the AI system; for this, data provenance is essential as it allows developers to understand data origins and any changes it may have undergone during its lifetime.

One problem that has received a lot of attention recently is bias, both in data and in algorithms. An example is the comprehensive study by Noble (2018) who describes in her book how search engines seem to have built-in biases against minorities. Another well-known example is Amazon's recruitment system (Reuters, 2018) had more men marked as successful on their recruitment database, which means the algorithm learned they preferred to appoint men. It isn't clear in this case whether the bias was down to the algorithm or the data, or both. Selecting data and getting it ready for processing (by adding labels, classes, etc) may reflect the biases of the people doing the work. These details have to be continuously fine-tuned. Getting data ready for processing can require a large amount of human intervention and pre-processing. The consensus is that it takes 80% of time to collect and clean the data and about 20% of time to analyse it (e.g. put it through machine learning processes).

2.3 Learning from data

At this stage, data structures and algorithms work together to make predictions using various models for processing data. As well as its role as input data for AI systems, data also plays a vital role in training, validation and testing AI outputs. The Data Governance WG's Data Governance Framework report outlines the need for different approaches to the governance of data, depending on whether it is used as training, validation and testing data (instrumental perspective), input or output data (data-specific perspective), and in the context of the wider data ecosystem.

At this step of AI development, data is used to create a test set and a training set. Training data is the set of data used for an AI system to understand how to apply and refine whatever techniques it employs to produce results. The quality and quantity of training data is important as any deficiencies present in training data may result in unreliable outcomes, decisions, or output data. Of particular focus from the perspective of data governance is the potential for any biases present in training data to result in the development of AI systems whose processes or outcomes may serve to reinforce these biases in the results it produces, be that novel data or information that feeds into human-led decision making. Validation and test data are used, respectively, to iteratively evaluate the AI system's operation (what it has learned from the training data), and to perform a final analysis on how well it performs its purpose, especially focused on the extent to which it is prepared to produce accurate results when 'real-world' data is introduced.

It is worth noting two aspects here:

- While the model training occurs, AI processes produce additional new data, potentially a large volume over time, which may need to be preserved (O'Leary, 2013). It also might be used for future training, and the decisions may influence the future behaviour of those it affects, creating a feedback cycle where previous behaviour of the AI system influences its future behaviour indirectly.
- Biases in training data are often identified at this stage and rectified by adjusting various parameters, sometimes by reprocessing data or by modifying the algorithm (Sun, 2019).

There is also pressure to build accurate predictive systems without using personal data, which may result in people using the wrong data and limit the amount of data available for training. Disaggregating certain data categories (e.g. gender, race) from processing may seem a good choice when building automatic decision making systems that perform tasks such as credit scoring tools, but including them may help identify patterns of discrimination in the first place.

Finally, the continuing access to data is of critical importance in developing AI. In systems that employ machine learning techniques to produce results, ensuring that the results produced are accurate requires an interactive process of training and retraining. However, changes in the governance or regulatory environment may result in data no longer being made available to systems that previously relied on it (for instance, data subjects may withdraw consent for their data to be processed, or diplomatic relations between states may result in data-sharing agreements to become redundant). This can potentially result in a deterioration of AI performance and, in the longer term, a lack of trust in AI systems to perform the actions that developers intend. Standardised practices and more formal regulations need to address the potential sensitivity or brittleness of AI systems and algorithmic processes in the event of access to data being withdrawn or data becoming unavailable.

2.4 Evaluation

Evaluation determines whether a system is ready for deployment or not. There are manual and automatic methods for evaluation. Manual methods include tasks performed by users, for example listening to an artificial voice and making a subjective judgement of how naturalness and intelligibility. Automatic methods include building statistics on various metrics, for example, Word Error Rate (WER)[3] in machine translation or speech recognition systems. Usability tests (task completion, surveys, focus groups) will also be carried out but these are mostly an assessment of the user-friendliness of the interface rather than of the underlying AI system.

One problem with assessing AI-based systems is that there are often no agreed industry standards on what constitutes good performance, for example, in natural language processing there is no agreement that a WER of 10% or less is good. Also, metrics don't always correlate with the users' assessment, they are often a reflection of the power of the algorithm. Manual and metric measurements can be complementary. For example, Toda et al (2004) made use of listeners' subjective evaluations to adjust cost-functions in the unit selection algorithm (in speech synthesis). Furthermore, while it is relatively easy to assess the product, it is not so easy to assess the underlying AI system. There are general principles such as explainability, fairness and transparency but factors such as commercial sensitivity will make it difficult to enforce them.

The 'National Artificial Intelligence R&D Strategic Plan: 2019 Update' (National Science and Technology Council, 2019) proposes a public-private collaboration to develop assessment criteria for AI systems based on standards and benchmark. Standards would govern the development of an AI system in terms of software engineering, performance, metrics, safety, usability, interoperability, security, privacy, traceability, and domain-specific standards. Benchmarks would be quantifiable measures of characteristics such as accuracy, complexity, trust and competency, risk and uncertainty, explainability, unintended bias, comparison to human performance, and economic impact.

2.5 Retention/preservation of data sets

Building data sets is very costly. It is therefore important to maximise their value by preserving them and as far as is possible, making them available for reuse. Making a data set available for further research and development activity may help keep it up to date as other researchers/developers are likely to contribute with new data. Data obsolescence is a problem in AI because accurate predictions cannot be made based on something which no longer represent reality. Wilkinson et al (2016) argue that other major drivers for reuse are cleanliness, accessibility and compliance with the FAIR[4] (Findable, Accessible, Interoperable and Reusable) principles. Retaining and continuously improving data helps improve AI models and for monetising, repurposing and recombining data assets that build up over time (which can give rise to new applications or value chains). In some cases, retention will be required for auditing purposes.

For data sets to be FAIR, good metadata (data about the data) is of key importance. In the first instance, for searching data, metadata should include the information needed so that the dataset can be discovered, fully understood and reused. For good quality data sets, data

[3] Word Error Rate (WER) is a metric used to measure performance of natural language processing systems. It measures the distance between the output sentence and a reference sentence at word level, by totalling the number of insertions, deletions, substitutions in the output sentence and dividing it by the total number of words in the reference sentence.

[4] See Wilkinson, M.D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship, Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

provenance, i.e., information about its origins and any changes (version history and time stamping) it has undergone since creation should be logged so that any new users can better assess the data and decide whether it is fit for purpose.

2.6 Deletion

Technically the training data is not required for the model to keep functioning so the data could be deleted. The possibility that the algorithm may need re-training and the costs of processing data means data is rarely discarded, and it keeps growing. Considering the resources needed to store the vast amount of data created, and the environmental impact of this (see more on environmental impact in section 5) appraising and selecting data for deletion should be considered an important step (see more on deletion in Section 6). There are various legal obligations related to the retention of data. An example of this is the European Union General Data Protection Regulation (GDPR 2020)[5], particularly article 17, Right to erasure (also known as 'the right to be forgotten'), which states that an individual can ask the data controller (the company that holds the data) to remove any personal records they may hold about them. The storage limitation principle, article 5(e) (GDPR 2020) states that data should not be kept for longer than necessary for the purposes of the project. It does not impose a time limit and it leaves it to the data controller to justify for how long they wish to retain the data.

Erasing personal data can be complicated as deleting records will require re-training of the model, which can take days and is expensive (and damaging to the environment). Ginart et al (2019) outline methods for deleting (encrypting and excluding records) for models employing k-means clustering. Veale et al (2018) discuss the relationship between regulations and intellectual property rights over AI models. It raises the question whether personal data which was used to train the model is the same as the data generated by the AI model based on the personal data. It also shows that, even after deletion of a record, cyber-attacks (model stealing via API, model inversion, membership inference) can estimate the training data and/or reveal whether an individual's records were part of the training. Deletion of data should be undertaken after appraisal and careful consideration.

Conclusion

This section has described the role that data plays at each step in a 'typical' AI development process. It has highlighted the importance of data and highlighted issues that AI developers must bear in mind when firstly selecting or creating data for the project at hand, how to iteratively evaluate the data along with the AI algorithms and models and ensure that sufficient data, and of good quality is used at every step. There are a variety of types of data used for AI development and specific criteria are used to assess their quality. Both types and characteristics of the data, will influence how it is/can be used in AI development and what actions need to be taken to ensure that data is fit for use for development of a specific AI technology. Section 3 will describe the key data types, while Section 4 will explain the concept of data quality and what characteristics data should have, to be considered as of good quality.

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1

3. Data Types Used in AI Development

Part of understanding the role that data plays in AI development involves looking closer at the types of data that are used by developers. This section will set out the main types of data available and the states that they appear in. We will also touch on how the availability of data types influences AI development, as well as looking at how the specific requirements of AI can play a role in the demand for certain types of data.

3.1 Data states

.Complex algorithmic and AI systems are required to process the vast amounts of data produced as a result of the increase of internet-based technologies in areas such as stock exchanges and financial services, industry and manufacturing, telecommunications and transport, and healthcare, as well as all areas of academia. In turn, these AI systems produce their own output data or new information in the form of categorisations or predictions. The data required to train AI systems in these areas, and the data that AI systems in these areas produce, are hugely varied in form. Before looking at the types of data used in the development of AI and their sources, it is worth considering first the states that data can be found in. We will look here at what is meant by structured and unstructured data and put these in the context of AI systems.

Structured data is data which is incorporated into a data model in order to standardise relations between data elements. In simpler terms, structured data is that which fits into a purposely designed, pre-defined structure. These models will generally have been designed with some particular goal in mind, for example, to store financial transaction receipts or to record the results of a controlled experiment, etc.; as such, structured data models can exist in many forms, from simple 2D spreadsheet arrays, to more complex relational databases or knowledge graphs. The key point is that the relationships between the elements in a structured dataset are defined by their position in relation to other data elements, with descriptions of the data and its meaning provided alongside the raw data (e.g. through metadata).

In terms of AI, this type of structured data is of use in (but not restricted to) supervised learning or in AI with limited and reactive capacities, which respond to specific input data. Supervised learning is designed to infer or determine the relations between pairs in input and output data elements. In order for supervised learning AI to be trained to an adequate level, it relies on the availability of high-quality data, with studies showing that structured data employing recognised standards is needed to maximise the value of the large volumes of data now available in areas such as medicine and healthcare (Pinto dos Santos and Baeßler 2018; Wang et al. 2020).

Unstructured data is data that is not organised according to any pre-existing data model. In general terms, what is considered ‘big data’ is unstructured data, at least in its raw form (see further discussion about big data in section 4.2 below). Unstructured data is unprocessed and is often generated by machine-led systems where the purpose of the data is not to answer a specific question; this includes, for example, social media posts, surveillance camera footage, or satellite imagery. As we can see from these examples, unstructured data can have its own internal structure, but what differentiates it from structured data is that the relationships between the data elements are often undefined. In addition, unstructured datasets require more (pre-)processing before they can be analysed or searched.

As the capabilities for gathering data has developed faster than the capabilities to analyse, more sophisticated AI systems are required to extract meaningful insight into unstructured data. Unsupervised learning is one technique used to gain insight in this area, whereby patterns and relations are identified in unlabelled and unstructured input data. This technique requires large amounts of training data and, though data used in unsupervised learning tends not to need as much human-led input in pre-processing and labelling, does require a 'human in the loop' at some stage(s) of development to test and verify the outputs.

The **labelling** (also known as annotation) of data is a key part of pre-processing data to prepare it for ingestion or for training an AI system which employs supervised learning. Labelling data involves assigning a meaningful tag to each data element in order that it can be identified and contextualised. For example, to use the study from dos Santos and Baeßler, chest cavity x-rays were labelled, denoting whether each example showed a definite case of lung collapse, probable case, no case, etc. to train an AI system to identify the condition (Pinto dos Santos and Baeßler 2018). Labelling data is a potentially time-consuming process and, when dealing with large volumes of data, often involves significant human input (as outlined in a report in the New York Times; Metz 2019), though in some cases AI systems can be trained to apply labels to structured datasets given sufficient examples.

Before moving on to look in more detail at data types used in AI development and their provenance, it is worth briefly considering how AI systems themselves play a role in the wider data ecosystem. In addressing the ways in which bias can manifest in AI systems, Ntoutsi et al. (2020) note that "algorithmic systems encourage the creation of very specific data collection infrastructures and policies". As has been set out above, different types of AI systems and machine learning techniques require different types of data. Algorithmic systems may rely on and require very specific types of data to function; added to this is the fact that access to certain types of data is easier than others. This creates a type of feedback loop where the available data determines which type of AI system can be used, with this AI system in turn limited in the type of data can be used as input data (especially in the case of supervised and limited machine learning). Recognition of the different types and states of data used in AI is necessary for accountability not only in explaining the results of AI-driven decision making, but in setting out ethical and legal governance policies for data in AI.

3.2 Data types

Understanding the various types of data, the specific features of each in the context of their applicability to AI development, and their sources can enhance transparency and work against potential 'black box effect' (Information Commissioner's Office (ICO) UK, 2017). This can also later help in any AI explanation tasks, in particular when it comes to (pre-)processing any data types, and should always be done in the knowledge that explanation of the AI model and its development may later be required.

As set out in the GPAI Data Governance Framework, data can be "classified into different technical categories of data according to a number of different criteria." We will look first at the main categories of data involving human actors in terms of their provenance (i.e. from where the data originates), in order to understand how each features in AI and how this impacts the longer-term availability of data for AI. Taking the lead from types of data set out in the Data Governance Framework report, the main types of data we will examine here are:

- Provided data
- Observed data
- Derived data

- Inferred data

In addition, we will look at two further types of data which relate to questions around data accessibility and governance. These are:

- Reference data
- Synthetic data

Provided data: Provided data refers to information provided by individuals, specifically those who are aware that they are actively providing data about themselves. The provision of this can be voluntary (for example, in the form of social media posts, financial transactions, personal emails, etc.) or individuals can be compelled to give their data (for example, in the form of registration forms for governmental organisations, health records, job applications, etc.). Individuals will be aware that their data is intended for use for specific purposes, with consent often required by data controllers. Its collection for use for specific purposes means that this type of data is more often found in a structured form, with labelled data elements. However, access to this type of data by those developing AI systems has its limitations; access regimes for personal data are generally restrictive due to the high degree of identifiability of personal data and the risks associated with this.

Observed data: Observed data consists of information gathered by observing actors or natural/technical phenomena in natural settings or environments. In research settings, the data generated by observational studies is collected with a view to using the sample observations to make general predictions or analyses of a wider population. In terms of observed data related to individuals outside of research settings, the degree to which an individual is aware of the collection of their data can vary. For certain activities, such as internet browsing or location activation on mobile devices, individuals may be aware that data related to these behaviours is being recorded. In other instances, individuals are less engaged or less aware that their behaviour is being observed and recorded in a digital form; examples of this include facial recognition software used in conjunction with CCTV footage or readings from sensor devices (movement sensors, light sensors, etc.). Depending on the context, observed data can be structured and unstructured; in addition, there can potentially be issues with data quality when contrasted with data generated in controlled research environments (Ross et. al. 2015). When involving data related to human actors, there are specific legal and ethical issues to be considered, especially in connection with the consent of individuals whose behaviours comprise a dataset which has been generated after a process or activity has occurred.

Derived data: This is data which is obtained by processing or applying some sort of a transformation to data that has been published or otherwise made available from any of the above sources. The types of processing or transformations include subsetting, changing structure, analysing, mining, or creating statistical or algorithmic models. Combinations of new and existing data sources and data types is one of areas where the value of data can be realised, not just in the context of AI but in all other areas (Frontier Technologies, 2020). However, this also potentially increases the ethical risks in terms of the use and misuse of personal data and the applications of data in areas beyond its intended original use.

Inferred data: This type data is generated by applying statistical or computational procedures to provided or observed data to produce data which can be used for predictive purposes. Though closely related to derived data, inferred data is more probabilistic in nature; derived data is more concerned with post-hoc pattern detection and categorisation (though it can be

used in later stages for prediction). Examples of inferred data, outlined by Abrams, include credit scores, likelihood of developing diseases, or creating targeted advertising (Abrams 2014). As some type of processing or analysis is carried out an original dataset to produce inferred data, there is a loss of control on behalf of individuals over how any personal data is used; this risk is outlined in the World Economic Forum report: "Understanding how the proportions of inferred and observed data are impacting the role of the individual is important to consider in policy formulation" (World Economic Forum, 2014).

Reference data: Reference data is used to give structure or to categorise other data or datasets, or to provide context for other data. Reference data can be either static or dynamic; examples of the former include fixed data objects which are constant, such as mathematical constants, lists of country code abbreviations, units of measurement. Examples of dynamic reference data include data which is variable but fixed in terms of relation to other data objects, such as opening and closing prices in financial markets or aggregated census records. Reference data is by definition typically highly structured in form and requires low levels of pre-processing to be incorporated into any procedures requiring data manipulation. Its value to AI development is in its combination with other data types or in providing cross-domain mappings for homogenous datasets, i.e. facilitating the combination of one or more other datasets.

As a sub-category of reference data important in enabling data sharing and reuse is metadata. Metadata is essentially information which provides the context for a given dataset. This can include information on provenance, data integrity tests, data formats, file size, etc. With respect to AI development, metadata is essential for the discoverability of datasets that can potentially be used in AI development. However, due to the specific requirements of AI systems, current protocols around the provision of metadata do not sufficiently address the applicability of datasets to AI development. For example, though a dataset's metadata may contain information regarding the type, volume and source of the data it refers to, it will not contain information on the steps taken to eliminate bias or whether the dataset has been disaggregated by race, ethnicity, gender, etc. Attempts to address barriers to the identification of relevant datasets for AI will depend on the sharing and searchability of metadata, where specific standards and protocols ought to be developed to encourage the documenting of AI-relevant characteristics of datasets.

Synthetic data: Synthetic data is all or in part artificially generated; that is, data that is not based on findings or observations based on real world phenomena, but on models and simulations of phenomena. Unlike the other data types outlined above, this data type is often initially produced by an AI or algorithmic system; it may also be produced by other methods, such as statistical or other data modelling that does not incorporate AI. This type of data can be used to inform decision making or predictions by human actors, or as a basis to inform further AI development, providing training or test data sets.

AI or algorithmic procedures can potentially be involved in each step of development of synthetic data, from development of the model to produce the synthetic data, analysis of this data, and the reuse of this data to train, test, evaluate or validate other AI systems. The accuracy or reliability of simulation data is dependent on prior knowledge of the system of phenomenon which is being simulated (Kim et al., 2017). This gives rise to the potential risk that simulated data, though 'correctly' corresponding to the predicted outcomes of the simulation model, may contain errors or biases which may then be reinforced if used further as training data or as input data for another system.

Conclusion

The increasing availability of all types of data is linked closely to the technological advances which has brought about the era of 'big data'. Though technological advances have also fundamentally altered how data is produced in others areas (for example, as in experimental medical and biological research, where computing capacity has resulted in techniques for whole genome sequencing), the technologies that have produced social networking platforms, cloud computing infrastructures, commercial transaction records databases, and the Internet of Things, to name a few, have resulted in the 'datafication' (Cukier and Meyer-Schoenburger, 2013) of all areas of human interaction. Though, as we have already mentioned, AI systems are dependent on the availability of data, there are ethical and legal ramifications of this proliferation of data which need to be addressed through proactive data governance, not least in the areas of data sovereignty and the rights of individuals over the use of their personal data. Issues with data quality and relevance, which will be explored in more detail in later sections, also play a role.

As noted by Cai and Zhu (Cai and Zhu, 2015) however, big data quality is low in comparison to other types of data outlined above, generally unstructured, with the variety of structures and data formats adding to potential difficulty in integrating data from various domains. Though big data and any datasets derived thereof are potentially of high value, the greater the reward for having AI intervene in activities (i.e. automated driving, disease detection, etc.) the greater the risk of negative impact if measures are not taken to address issues resulting from substandard data governance protocols or the potential ethical pitfalls with different types of data.

4. Data Characteristics that Influence the Process or Outcome of AI Development

This section outlines data characteristics that influence AI development and link to recent research and development from a range of fields. We will attempt to define data quality, outline the characteristics that comprise quality data in the context of AI development, with a specific focus on how data governance practices have a role to play in creating data that drives responsible AI.

4.1 Data quality and data governance

Quality data is crucial in all stages of AI development, and while certain types of AI are designed to deal with unstructured or low quality input data (e.g., due to size, timeliness, biased data and other factors as this section will outline), the quality of the data used in the development stages is of importance in order to reach the point of AI making accurate, valid and unbiased real world decisions. As of yet, there is no single definition of data quality and multiple approaches have been made in defining benchmarks for data quality (e.g., The Taxonomy of Dirty Data by Wong Kim et al. (2013) and Wakchaure et al. (2008) on algorithms measuring data quality for a criminal records database).

There is however a consensus that the common denominator for quality data is 'data that is fit for use', meets specifications, requirements and expectations (Fürber, 2015). The literature on the topic of data quality groups characteristics by similarity or common attributes. In this regard, Wang and Strong (1996) define four categories of data characteristics, listed as 'intrinsic, contextual, representational and accessibility related'. Batini and Scannapieco (2016) define data characteristics in 'clusters' – the accuracy, completeness, accessibility, consistency and readability clusters. For the purposes of this report, the focus is on combining elements of the two dimensions and clustering approaches referenced above where there is relevance for AI development, while expanding to include FAIR data as well as sensitive data and the care to avoid bias. The following table summarises the characteristics that will be examined in this section.

Grouping	Characteristics
Sensitivity	Inclusion of protected characteristics and causal influences, identifiability of individuals, anonymisation, commercially sensitive data.
Representativeness	Bias in data collection (selection bias, exclusion bias, reporting bias, detection bias) leading to over or under representation
Accuracy	Objectivity, precision, reliability, validity, legitimacy, labelled data
Completeness	Timeliness, appropriateness, volume of data
Accessibility	FAIR data, open or closed data, Level of standardisation/ interoperability, access, security
Coverage	Geographical and temporal coverage, representational consistency

The use and management of sensitive data may refer to **personal** or **non-personal** data. In the case of personal data, an important factor is the inclusion of **protected characteristics,** as

defined by the UN[6]. This term refers to 'sensitive attributes' such as gender, sexual orientation and identity, minorities and indigenous people, religion and disabilities, which may lead to bias or unfair discrimination (Silberg and Manyika, 2019). There are a number of considerations relevant to sensitive data: inclusion and correct level of representation (further explored in section 5.3), bias elimination through both data management and algorithm development as well as the handling and security of sensitive data, which will be further examined in Section 7.

Simply removing sensitive characteristics from datasets is not the solution to eliminating discrimination as algorithms can find proxy indicators to convey the bias; for example, values relating to race might be removed, but an individual's post code already be associated with race (Bruno et al., 2017; Ntoutsi et al., 2020). An important step to bias elimination is being aware of such inferences due to 'redundant encodings' and locating and understanding these **causal influences** (Ntoutsi et al., 2020). Ultimately, an algorithm should be able to reach the same decision whether values related to race or gender are removed from the equation (and causal influences are not affecting the decision-making process). An exemption of this would be in cases of affirmative action, where fair decision making processes would require special attention given to minority communities (for example, in cases of government contracting where the intent is to assist those regularly affected by bias and discrimination - while still bearing in mind that this practice should not have significant repercussions for others (Xiang and Ho, 2020).

Sensitive data can also refer to **commercially sensitive data,** defined as data of economic value that would be damaging to, for example, a business or other commercial entity if released (Rosenblum and Maples, 2009). This type of data is generally considered 'less sensitive' compared to personal data. However, it should be noted that from an AI development perspective, while bearing in mind the need for transparent AI, disclosing details on how the AI works might create vulnerabilities for AI developers and providers to their own commercially sensitive data - i.e., expose how the software works (ICO, 2020). Generally, complying with data protection regulations does not require disclosing sensitive information such as software code.

4.3 Representativeness

For an algorithm to work, the training data needs to be representative of the real-world demographic on which it will be used. Therefore, the first issue is **underrepresentation** of protected characteristics. A lot of discussion has been generated around bias in facial recognition and gender classification tools with the literature indicating varying degrees of algorithm accuracy for different demographic groups. Phillips et al. (2003) report that younger subjects and white males are more likely to be accurately recognised. A study by NIST using four datasets from different sources in the USA (justice system, immigration and border control) attempted to quantify the accuracy of facial recognition algorithms defined by sex, age, and race or country of birth (Grother, et al., 2019). The results indicated high rates of false positives (i.e. where faces of subjects were similar, there was false association of samples) in female, indigenous peoples, Asians and African American subjects. In contrast, where the algorithms were developed in China, the false positives on images of Asian subjects were significantly lower. This difference in algorithmic performance based on development location further stresses the importance of **representativeness** in training data used. Simply put, if the training and validation data is not representative of the real-world population, AI is

[6] Please refer to the UN guidance of protected characteristics for a comprehensive list: https://www.un.org/ruleoflaw/thematic-areas/human-rights/equality-and-non-discrimination/

likely to fail to recognise them in real word applications, which risks amplifying existing inequalities.

Overrepresentation of sensitive characteristics is the other side of this issue where bias can be introduced through the data collection and selection process. Criminal justice models and predictive policing algorithms are frequently used examples that illustrate the use of biased data. This is because specific neighbourhoods with high correlations for race are selected for training algorithms sampling - areas that are already known to the police officers - leading to further increase in policing in those areas, new crimes being observed and thus higher crime statistics as well as even higher policing levels (Silberg and Manyik, 2019; Lum and Isaac, 2016). Bias in policing thus finds its way into the data and is further replicated and amplified in AI development in this field. It is worth noting that at the European level, the European Commission has made policy recommendations for use of data sets with broader representation in terms of gender and ethnicity (European Commission, 2020). However, in some instances the protected characteristics are not available to the organisation that needs to validate data to eliminate bias and in some jurisdictions it is illegal to collect data from specifically targeted groups, and it may be illegal to collect information on protected characteristics in order to assess the presence of bias in a dataset or algorithm.

Another issue pertaining to sensitive data is that of **anonymity** and the **identifiability** of individuals when using or having a need to publish personal data. Preserving privacy has been a prevalent issue with the field looking at approaches based on generalisation – replacing personal values with less specific but accurate alternatives, in order to avoid the potential for linking neighboring quantifiers to specific individuals (Samarati and Sweeney, 1998). This endeavour extends beyond directly personal data (e.g. name, address) but also data that can reveal behaviours of individuals, such as movement data from mobile phone and GPS applications (Andrienko et al., 2009). Other privacy preserving endeavours are centred around the AI integration with Blockchain (Panda and Jena, 2020). Machine Learning models using medical data have examined anonymisation and differential privacy, introducing noise in a dataset while preserving outcomes (Gaur, 2020). COVID-19 has created the need for sharing data – for example, patient imaging data – while ensuring identifying characteristics are removed prior to sharing and processing (Ulhaq and Burnmeister, 2020) in order to conform to, for example, GDPR which poses limitations on global sharing of personal data.

From a governance perspective, transparency in input, output and source code is of utmost importance (where possible); specifics of these concepts in practice can be more difficult to determine, especially where there are a number of stakeholders involved in data creation and use. Also, of importance is adopting auditing strategies (e.g., decision process as a black box), and having clarity on where the decision making and accountability lies (Lepri et al., 2017). Privacy risk mitigation when sharing data, as well as informing users of risks should be the responsibility of AI-based system developers and service providers (Findlater et al., 2020).

4.4 Accuracy

Accuracy of data refers to the **reliability** of information, the assumption that the information or value conveys the true state of the source, is factually correct and unambiguous (Cai and Zhu, 2015), while **precision** is the 'closeness of agreement between test results' (International Organisation for Standardisation, 1998). Accuracy of a given data value is measured by comparing to a known reference value. The degree to which accuracy can be ascertained varies and can depend on context or additional information in order to be verified. Curating data and evaluating accuracy often requires input from trained experts in the field as well as data curators. However, in many instances verification and an assessment of precision is not possible, e.g., in the case of data corpuses consisting of social media posts.

The importance of data accuracy can vary depending on the field of research, application and the degree to which accuracy can be verified among other factors. Firstly, accuracy is often not easily discernible, and further data sources may be needed to supplement it – as is the case with social media data, where **credibility** serves as an additional qualifier (Cai and Zhu, 2015). The intended outcome also defines whether accuracy is high priority or even of benefit. In an HR setting, where the system's function is to detect bias in previous recruitment decisions, a dataset used for training consisting of accurate data is required. Where the intended outcome is a fair recruitment process, using this dataset of historical decisions for AI training could result in biased decisions (see: GPAI Data Governance Framework).

Objectivity and **credibility of data** can also depend on whether data is raw or interpreted. An example from the medical field where raw data is encoded in order to translate into billing codes indicates that despite employing trained coding staff, interpreting medical staff notes and diagnoses creates a layer of human judgement and as a consequence a degree of subjectiveness (Strong et al, 1997). **Labelling** data can also introduce a degree of subjectiveness; particularly if this task is carried out by an internal team with specific expectations about the outcome, bias can sneak into the process (Pang, 2019).
To ensure accuracy in a dataset, the following steps can be taken: 1) benchmarking, 2) comparative analysis of consistency, and 3) auditing. These steps can be demonstrated in practice using the example of a team of phonetic transcribers annotating large audio-visual datasets for an Automatic Speech Recognition system, as this is an effort intensive, manual task involving a degree of human judgement with potential for error.

Beyond data accuracy, the quality of a dataset also depends on the inclusion of valid data. **Validity** refers to the extent to which the data reflects the real world. In the context of AI, data used in development should adequately reflect the reality of the real-world situations on which it will be applied to (Son, 2020). Based on the assumption that validity can be confirmed by connecting data elicitation approaches to theoretical constructs (Howison, Wiggins and Crowston, 2011), the validity of a dataset used for AI training can be assessed by using this training dataset as test data for a different AI system, and observing the results.

4.5 Completeness

Data **completeness** refers to data with no missing values; a complete dataset has no deficiencies that affect the use of the data, or impact data accuracy and integrity (Cai and Zhu, 2015). Incompleteness is a common characteristic of low-quality data, and it can be addressed by data fixing or imputation, which can often be very effort intensive. Determining the completeness of a dataset provides valuable input on whether it is suitable for querying, mining and analysing (Liu and Zhu 2016). It is important to note that there should be transparency where a dataset has been reconstructed using algorithms or delivered in a complete state.

Timeliness is defined as the time between data being expected and becoming available and accessible for use (Loshin, 2009), or data that represent the required point in time (Sebastian-Coleman, 2013). One of the challenges due to fast changes in big data is the fact that if data is not acquired in real time or dealt with in sufficient time, there is a risk of using out of date information and as a consequence producing inaccurate results and making wrong decisions/predictions with potential harmful financial and ethical implications (Cai and Zhu, 2015). Therefore, since an algorithm making predictions based on temporal data needs timely data in order to make accurate predictions for the future, it is expected that for such applications training data will need to be updated and systems retrained with more up to date information (depending on the system and scope of the project) as needed (Perlin, 2020).

For an AI system to make accurate predictions, the dataset needs to include **relevant** data. An AI system predicting future behaviour of a company's customer base with regards to sales in the UK would require information on recent customer behaviour of this specific market segment; a dataset with information on the US customer base – no matter how accurate, complete or clean would not yield the intended results for this specific demographic.

4.6 Accessibility

Once relevant datasets have been identified, **accessibility,** i.e. whether specific data can be accessed or purchased, is an important consideration, as AI development thrives on access to big and varied datasets. Accessibility directly relates to a number of other factors:

a. **Legislation** Going back to section 5.1 above, sensitive data might be inaccessible and protected under privacy regulations such as the GDPR or HIPAA, so access can range from closed, to shared to open.
b. **Legal and administrative barriers** can delay data accessibility, which affects **timeliness** as per section 4.4 above. Data needs to be accessible at the point in time when it is needed, this is especially important in fields such as pandemic research and disaster response to assist with the development of AI technologies that are a part of a rapid swift response.
c. **FAIRness of data**, and ensuring that 'the human or machine is provided - through metadata - with the precise conditions by which the data are accessible, and that the mechanisms and technical protocols for data access are implemented such that the data and/or metadata can be accessed and used at scale, by machines, across the web' European Commission, 2018:19).
d. **Findability and discoverability**. Although data is openly available and accessible, it may be hard to find, for example if it is simply stored on a website which is not indexed as such. Metadata (data about the data), data citation, use of repositories and assignment of DOIs to datasets are important in ensuring that datasets can be found and discovered by those who need them. This will also allow for the tracking of use, which will assist with data provenance (in cases where datasets are altered) and in cases where data may need to be redacted due to bias or errors.
e. **Commercial restrictions**. In many instances data is inaccessible due to commercial restrictions and ownership. Currently there are several corporate data owners (e.g., Microsoft, Google and Amazon) which hold vast amounts of data that are largely inaccessible to AI developers.
f. **Licensing Data** may be available and accessible, but terms of its use may be unclear, due to lack of licensing. Data may also be unusable due to strict licensing. Recommendations in this respect that as far as is possible, accessible data should have clear licensing terms, which clearly state conditions for its reuse.

Making data accessible requires effort and funding so that data can be firstly made into a reusable resource and secondly can be curated to securely stored and easily located in. This does present a barrier to data work and accessibility, for example in low- and middle-income countries where funding for such development is limited[7].

A highly relevant and related note on the discussion of openness and accessibility of data is one on data colonialism, which centres on the rights of communities and nations to work on

[7] A recent initiative, Lacuna Fund was launched in 2019 to fill data gaps in Language, Agriculture and Health datasets that will enable further development of AI in the specific fields of Communication, Health and Agriculture. https://lacunafund.org/

and benefit from their data. As stated above, low and middle-income countries may lack resources for data work and there have been instances of organisations from higher income countries undertaking data collection or use of data from communities, without any benefits to the local populations. Also, there are well known instances of non-reciprocal sharing of data and knowledge, for example in the response to the Ebola virus whereby data from the treatments of Western patients was not shared with scientists working in Africa. At the same time data collection was being undertaken in many countries affected by the virus and access was not given to local scientists. (WHO, 2015)

Here it is also important to note the Indigenous Data Sovereignty, as 'the right of Indigenous Peoples to own, control, access and possess data that derive from them, and which pertain to their members, knowledge systems, customs or territories' (IWGIA, 2020), should be considered in any discussions about accessibility, openness and FAIRness in relevance to data from indigenous populations. The rights of indigenous communities to create and derive value from their own data should be respected and data should be used in ways which align with indigenous values and worldviews. The CARE principles are an important tool which highlight these issues, and were written to complement the FAIR principles, to ensure that indigenous data rights are considered at all stages of the lifecycle of indigenous data. (Global Indigenous Data Alliance: 2018)

4.7 Coverage

Coverage can refer to the representativeness of a dataset in terms of geographical and demographic representation, or of enough representative samples as required by the specifics of each project. High coverage in a dataset is important to avoid bias (please also refer to sections 5.1 and 8). Depending on other parameters, project scope and the question being addressed, some geographical bias might be accepted and too wide a site sample might be detrimental (Field et al, 2017 provide an example from research on species behaviours across various sites). In Geospatial AI development, it can refer to 'digital geospatial information representing space/time-varying phenomena' (Open Geospatial Consortium, 2012), as well as geographic coverage, i.e. coverage across the globe, and temporal coverage, i.e. 'frequent revisit times' (VoPham et al, 2017). Good data coverage might be challenging in a healthcare setting, in low and middle-income countries where access to healthcare or representative data is not available, and in high income countries where access to healthcare may be unequal due to costs.

Conclusion

This section has described the characteristics of data inherent in data quality and provided examples of how data of poor quality can impact on the resulting AI technology. Specific processes exist that can help assess data so that any quality issues can be mitigated, such as Data Quality Assessments and Data Gap Analysis.

Conducting a Data Quality Assessment (DQA) at the beginning of a project is the first step for outlining the strategic goals, contextual requirements, as well as deciding what types of data are needed to achieve the project aims. This will include prioritisation of the importance of quality criteria and data characteristics should be undertaken and planned accordingly (Cai and Zhu, 2015). Furthermore, performing a Data Gap Analysis before embarking on a new project assists to clarify the requirements and needs of the project, examine and assess the quality of existing data and how it compares with the desired state, compare against the objectives and shed light into potential issues early on.

Ensuring data quality throughout the AI development process through good Data Governance is an important component in avoiding what has been simply described as 'garbage in, garbage out', where poor quality data leads to poor results.

The impact of data on AI results, specifically regarding socio-ethical, economic and environmental issues will be explored further in the next section. The focus will also be on highlighting the effects of law and transparency on data as well as exploring challenges arising from lack of access to specific data.

5. Socio-Ethical, Economic and Environmental Impact of Data in AI

Data-related issues emerging from the role of data at each step of AI development are indeed very complex to be extensively analysed here and out of the scope of this report. However, this section offers a wide mapping of important issues emerging from the analysis and discussion so far in the report on the data types and characteristics, their accessibility and availability in the AI development. The section provides a brief examination of benefits as well as potential harms on socio-ethical, economic and environmental levels feeding directly to recommendations for the future work of the GPAI WG in Data Governance and inform the future developments of AI.

5.1 Socio-ethical impact

Data-driven AI raises both challenges and opportunities related to socio-ethical impacts, including but not limited to the following concepts: do no harm (non-maleficence) and do good (beneficence); trust; dignity; inclusion and exclusion; privacy; asymmetries of power between users and service providers; and equity of opportunity to access services. Numerous frameworks and texts outline ethical principles for responsible AI (such as OECD 2019b; EC HLEG Ethics guidelines for trustworthy AI 2019; EC European Ethical Charter on the use of AI). It can also enable or threaten certain human rights. For instance, data-driven AI can ultimately have a positive impact on financial services, manufacturing, healthcare, governmental services and research to name a few. These positive impacts may include, for instance, finding previously unknown, important correlations between data sets; using this to better distribute resources; and to encourage collaboration that leads to better quality data. There can, however, be negative effects of accessing and using data. It could create new forms of vulnerability by perpetuating and reinforcing inequalities on macro and micro levels. This section explores the socio-ethical ramifications of data-driven AI, with a specific focus on manifestations of autonomy, beneficence, non-maleficence and justice; privacy; inclusion and exclusion; and equality and non-discrimination.

Autonomy, beneficence, non-maleficence and justice

The healthcare sector is rapidly increasing its use of AI. This subsection will therefore focus on the healthcare sector when describing some of the socio-ethical effects which may emerge during the development of AI. There are four key principles of healthcare ethics: autonomy, beneficence, non-maleficence and justice (Beauchamp & Childress 1985). Medical professionals accept them as valid and action-guiding. Using data to enable the development of AI in healthcare has led to advantages for both patients and healthcare providers. The use of big data in healthcare is increasingly prevalent and there are multiple factors that are driving the necessary growth in this field.

There are numerous concrete examples of how data-driven AI can improve healthcare by developing innovative solutions to support decision-making and improve diagnosis and general efficiency of the healthcare system (EC HLEG 2019). It can improve patient experience by providing care robots, virtual assistants and predictive applications. AI may also contribute to overall efficiency of the healthcare sector by combining datasets to identify wasted resources or inconsistencies in, for instance, individual hospitals or even national healthcare systems (Lomas 2018).

Nonetheless, the use of AI in healthcare has presented serious socio-ethical issues. In a healthcare context, there is a clear contrast between the abundant, detailed datasets that data brokers own, and the scant, disconnected datasets within the sector (Fry 2018). Data can exist, but it can be recorded in different forms, which makes it difficult for the data to be

useful (ibid). The issues within this sector mainly revolve around the data that is used to train certain algorithms that then, in turn, issue outputs that may be biased and thus, affect individuals and communities. Missing data, sample size, and misclassification or measurement error (Gianfrancesco et al. 2018) may all lead to bias. Often, these issues are related to data access, however, practitioners could also translate their implicit biases, for instance those related to gender, into documentation that becomes input data for these algorithms. In the health sector, this process could then negatively affect the quality of care for underrepresented or marginalised groups towards whom a practitioner exhibits his or her implicit bias. One such issue is potential error at scale associated with the automation, as explored below. See also section 7 for examples of use of AI in pandemic response.

Example 1: Healthcare and unrepresentative data sets that threaten justice

In healthcare ethics, the principle of justice denotes an element of fairness in all medical decisions. Issues can arise with missing data. For instance, certain data of patients who have visited multiple healthcare facilities or have been able to access only online portals could be missing from training data sets, thus leading to inaccuracies in predictions related to these groups (Gianfrancesco et al. 2018, p.1545). Also, small-scale patients' populations could be omitted from a sample set, thus excluding them from clinical decisions based on these algorithms (ibid). This would hamper the guiding notion of justice, as the resources and treatments would not be equally distributed and there would be an element of unfairness.

Example 2: How the misclassification of data in healthcare could threaten non-maleficence and justice

The principles of non-maleficence (do no harm) and justice could be threatened by using data-driven AI in the healthcare sector. Several factors could influence the misclassification of data. For instance, patients of lower socioeconomic status or uninsured patients could be more likely to visit teaching clinics rather than, for instance, practices or hospitals where patients of higher socioeconomic status or with insurance are more common (ibid, p.1546). The former situation could lead to less accurate documentation or diagnoses, thus contributing unfairly to errors related to minority groups of patients. Furthermore, inaccurate diagnoses could lead ultimately to a medical practitioner harming the patient or groups of individuals.

Privacy

As the process of developing AI becomes increasingly sophisticated, the analysis of people's personal data or information can lead to heightened privacy concerns about the information pertaining to them. The types of data and how developers use them is a key consideration. Recent scholarship has identified privacy concerns arising with the use of data-driven AI in the contexts of, amongst others, law enforcement (Rowe and Muir 2019), advertising (Estrada-Jiménez et al. 2019) and border security (Beduschi 2020). Such concerns include the collection of sensitive personal data that is then combined with other datasets to build a profile about a person without their knowledge; this profile is then used to target them in a security context, which could be efficient and beneficial, or inaccurate and unfair.

The effectiveness and fairness of AI-generated output decisions are also dependent on good quality input data about someone's earnings, criminal record, or social care and education.

Public institutions can be pressured to repurpose personal data which has been previously collected for a specific purpose. This could violate the EU General Data Protection Regulation (GDPR) obligations of purpose limitation and limits on the secondary use of personal data, thereby threatening an individual's data protection and privacy rights (Choroszewicz and Mäihäniemi 2020). Moreover, if an algorithm then bases its automated decisions on low-quality or unrepresentative data, these systems can produce identical, problematic biases or unfair discrimination at an even greater scale than those produced by human decisions, creating larger problems for a broader public (ibid). Section 7.1 discusses privacy and data protection concerns in the field of pandemic response and how these have manifested in AI development for COVID-19 diagnosis and drug repurposing.

Inclusion and exclusion

Inclusion is about ensuring all humans can participate in an activity as valued, respected and contributing members of society; it requires proactive, positive action. On the other hand, exclusion can be understood as a dominant group considering their own norms superior, and thus marginalising or disparaging others outside of this group. It is important for data-driven AI to ensure inclusion and avoid exclusion.

In terms of health data, there are multiple issues that can have worrisome ramifications. Indeed, it is a field 'where data are notoriously messy' (OECD 2020a). Data are unstandardised and reflect specific patient populations; biased decision-makers can then make errors that the resulting output data goes on to reflect (OECD 2020a, p.3). For instance, certain disease predictions made by an AI model can lead to unreliable outcomes that need extra verification by clinicians, which in turn threatens their trust in the process. Such predictions, based on input data about a specific population (see section 4.3 on data representativeness), could then fail to be applicable to different populations, which could consequently perpetuate existing inequalities and biases.

The public sector and its use of AI may further illustrate this point. Governments worldwide are increasingly using data and new technology, including AI, to automate their processes and procedures. Research has shown, however, that in this context, human rights of the poorest and most vulnerable people are at greater risk. Various governments are using similar data to re-engineer social services to move towards a 'digital welfare state' (Lomas 2018). For example, public sector institutions in, amongst others, Denmark, Sweden, Germany, Finland, France and the Netherlands use automated decision-making on welfare entitlement, fraud detection, criminal risk assessment and child protection services based on citizen data (Choroszewicz and Mäihäniemi 2020). Similarly, in the UK, citizens' data are used on local and national level for more efficient governance combining data from a variety of sources (Dencik, Hintz, Redden, and Warne 2018, p.3).

As goods and services are increasingly related to AI, issues surrounding the ethical questions of inclusion and exclusion can surface when examining the data used to inform AI. These concerns can sometimes stem from historical bias that is then aligned with new methods. For instance, as discussed in the context of AI and healthcare, when there is an absence of data on some populations, misalignment may occur and thus poor decision-making. A lack of labelled data or poor-quality metadata could amplify already-existing vulnerabilities and inequalities; there could also be discrepancies in someone's options to refuse data collection.

In a governmental services context, the UN Special Rapporteur on extreme poverty and human rights has stated that digitised welfare states would have an immense impact on vulnerable people (Alston A/74/48037 2019). Moreover, referring to a digital welfare state approach in the UK, he has noted several data-related issues (ibid). These issues can surface

in the development of AI for multiple purposes, as illustrated with the two examples below. It is easier to obtain or use data about some people, groups and geographical regions than others, which can also lead to issues of exclusion.

Example 3: social security payments and algorithm design for automated decision-making

In Australia, roughly 500,000 mistaken debt notices were sent to social security recipients because the partly-automated system averaged earnings over a series of fortnights rather than actual earnings in one fortnight leading to errors that affected a particularly vulnerable proportion of the population (Carney 2019).

This exemplifies how a data feed could include inaccurate, outdated or overdue information that would unfairly affect vulnerable people about whom decisions are being made automatically. In sum, socio-ethical issues pertaining to bias can surface throughout the process of developing AI. As such, such systems would not enable inclusion and could unfairly exclude members of society.

Equality and non-discrimination

The concepts of equality and non-discrimination are inextricably linked. Equality means sameness and equivalence in relevant respects, such as value (IEEE 2017). People are equal in terms of their human rights, status and opportunities. Discrimination is when an entity (technological or otherwise) differentiates between categories of persons to determine their entitlements, rights or eligibility (ibid). Discrimination becomes a problem if someone is treated unfavourably because of this differentiation. Indeed, the principle of non-discrimination seeks to guarantee human rights to everyone, without discrimination based on race, nationality, gender, language, religion, ability, age and similar. AI must be designed and deployed in a way that does not threaten quality or violate principles of non-discrimination. Below are two examples of AI-enabled threats to the principle of non-discrimination.

Racial discrimination

In terms of diversity, most AI tools are being researched and developed in regions (specifically, the US, EU, UK, Canada and Australia) that train their models on input data pertaining to populations in Western, industrialized and democratic countries (OECD.AI 2020). As demonstrated by IBM's venture into AI for healthcare, it is very difficult to compare one patient (from one population) to previous patients (potentially from different populations) to make an informed, reliable and unbiased diagnosis or decision (Strickland 2019).

Example 4: healthcare and lack of data context enabling racial discrimination

Beyond population or geographical diversity, there exist issues with racial discrimination in data-driven AI for healthcare as shown for a popular commercial algorithm used in the US healthcare industry (Obermeyer et al. 2019). In particular, the study showed that it wrongly assigned Black patients the same risk level as white patients leading to lack of appropriate care. It had to do with (real) data reflecting lower costs of care for Black Americans – because of access and payment differences. As such, an algorithm that equated cost with illness severity, underestimated Black peoples' need or qualification for healthcare programmes. In other words, this reflected a lack of understanding of context and data generation, and not a problem with the data, which reflected reality.

Example 5: hate speech, racial bias and mis-labelled data

A study showed unexpected correlations between markers of African American English (AAE) and rates of toxic speech in widely used hate speech datasets (Sap et al. 2019). The problem was that Tweets in AAE or by African Americans were up to twice as likely to be labelled erroneously as offensive due to being trained on models that propagated these biases (ibid).

Gender discrimination

Research suggests that gaps in big data influence gender politics; a gender data gap has plausibly silenced women and erased some of their accomplishments, narratives and experiences (Criado Perez 2019). Algorithms can perpetuate negative biases against, for example, women of color by embedding them in search engine results or hate speech detection methods (Noble 2018). Amazon's biased hiring algorithm is a simple example to illustrate the amplification through AI of existing gender biases in the society inherited in the data and therefore in the algorithms.

Conclusion

When considering data governance, relevant stakeholders ought to consider the socio-ethical opportunities and challenges that various data-driven AI raises. Throughout the AI development process, ethical issues related to data types and data characteristics are evident. Tangible ethical issues explored here include healthcare ethics, questions of privacy and data protection, inclusion and exclusion, and equality and non-discrimination.

5.3 Economic impact

Introduction

Data-driven technologies aim to offer cost efficiency in critical sectors such as governance and healthcare as discussed above. Thus, data is often presented as having inherent potential economic value, as captured by the much-used phrase 'data is the new oil'. At the same time, the "new data revolution" raises concerns on the changes of work and work-relationships as we know them resulting in an enormous economic outcome (Ishmail, 2018). This section discusses the advantages of the increasing data-driven technologies as well as the potential negative impact they can have on national and global economies.

Economic advantage

The economic advantages that could arise from using data-driven decision making are considerable. Data can be used in AI to increase efficiency and reduce costs by automating otherwise time and resource-consuming activities, for instance Big Data analytics is able to consider large amounts of information near instantaneously (Alam et al. 2014, p.446). Moreover, the depth of insights can be improved by utilising an array of data points to provide a more nuanced understanding and develop knowledge that was not previously attainable, for example through the collection of behavioural information (Huyer and Knippenberg, 2020, p.62). Moreover, there is an economic benefit in the development of new technology. Data can be used to improve products and tools, as demonstrated through the increased number of 'smart' products that are now on or entering the market (Nunes, Pereira, & Alves 2017, p.1218). The expanded availability of data has led to a relative boom, creating a considerable number of jobs in this area and powerful innovation hubs or regions such as Silicon Valley in California or Estonia.

Surveillance Capitalism

These developments are, ultimately, dependent on data. How data is sourced and used is of central importance when considering the economic impacts. Some of this data is produced through arguably uncontroversial sources and utilised for the public good, such as the Humanitarian Data Exchange platform operated by the United Nations Office for the Coordination of Humanitarian Affairs (Zwijneneburg et al. 2020, p.355). However, in some instances data collection has acted as a business model. As data is increasingly harvested and processed for economic purposes, it has a direct bearing on how economies are shaped. Perhaps the phenomenon of 'surveillance capitalism' best encapsulates this. Under this phenomenon, companies provide services to users, often for free, whilst harvesting the same user's data in order to study behaviours and utilise this information for financial gain as for example through targeted advertisements (Zuboff 2015, p.79). Major companies using data as their business model have extracted huge profits (Zuboff 2019).

The rampant success of this business model has allowed companies (operating with little regulation) to monopolise these data sources (Chandran 2020). As successful social media platforms develop, they are often bought by the major players within the market (Levy 2020). While this can provide an incentive for financial reward for those seeking large payouts, it can also act to disrupt the market and stifle innovation from smaller organisations. The costs of data access from these monopolies can create inequalities and has become a burden for the creation of AI for smaller stakeholders (Carriere-Swallow and Haksar 2019, p.32). As such, increased anti-trust restrictions have been offered to allow greater competition and break up the relative stranglehold that larger companies have over user data (BBC News 2020). Despite this, the data gathered from these platforms has been utilised for public good as well. For example, social media crawling has been used to develop real-time awareness of natural disasters (Joseph et al. 2018, p.287).

Economic Impact of Regulation

As countries and regional blocs have taken steps to approach data through a rights-based framework, regulations have been developed, such as the GDPR within the EU (Perera et al 2019, p.404). Given the scale of data processing within society, these regulatory developments have had a monumental effect. This not only relates to the fines that stem from grave violations which can amount to €20m or 4% of annual turnover, but much more significantly to the compliance costs as per article 83(5).

In the run up to the GDPR coming into force, companies rushed to understand the compliance requirements, amend their current practices, and develop the necessary procedures to ensure future compliance. In 2018, Veritas reported that on average, firms were forecasting spending in excess of €1.3m ($1.4m) on GDPR readiness initiatives (Veritas 2017). Such costs continue following the implementation of the regulation in national legal systems. In this respect, ensuring that they have the requisite knowledge, capacity and capabilities creates a significant economic impact.

The resource requirements necessary to comply with the GDPR have resulted in an unlevel playing field. Whilst simple GDPR explainers are easy to find and access online, smaller businesses and organisations are less able to dedicate the resources for compliance that larger companies are able to. Smaller businesses may in fact, lack the dedicated personnel or ability to outsource (BBC News 2018). The imbalance in the ability to respond to the new regulatory requirements may lead to larger organisations developing a monopoly over services at the expense of smaller organisations who do not have the required budgets and inhouse expertise.

Economic Marginalisation

The societal impacts of the unethical use of data can be severe and can lead to the continual exclusion of certain persons from opportunity. This marginalisation could have a wider effect on the economy by entrenching inequality and offering advantage to the more affluent. The sheer volume of data available by monitoring online behaviours offers companies immediate access to data points that were not previously available. This can allow loan companies, for example, to incorporate seemingly arbitrary details such as the way an email address is formulated into their determinations on whether to grant or deny credit (Klein 2019). The increasing availability of data could lead to further marginalisation and perpetuate economic disadvantage. This result could deny less-affluent communities from benefiting from economic opportunities and minimising the weight of their participation as poorer communities are often less captured in data sets (Cinnamon 2017, p.617).

Conclusion

Ultimately, data can have a considerable beneficial effect for economies through increased efficiency and depth of insight. Moreover, its role in innovation can boost the economy through the development of new products and solutions (for example within the health sector, professional, scientific or technical services). However, the underlying structural issues indicate that there is a potential for data to increase inequality, long-term economic harm through its misuse. Indeed, if mishandled, data could lead to the exclusion of certain individuals from financial rewards, and the concentration of data in the hands of major companies which may stifle innovation practices. Poor regulation of industry, labour standards, access to training and education, and social welfare are all inextricably linked to sustainable growth and economic stability. Thus, these issues must be properly addressed and appreciated through policy and data must not be deployed at the expense of the public good.

5.4 Environmental impact

The impact of the ever-increasing data collection, process and storage for AI development on the environment is often under-discussed in public discourse. Since the creation of the SDGs, the need to monitor their development has led to a reliance on data and an investment in efforts to improve data monitoring (United Nations, 2015). Data can be utilised to better protect the environment and promote sustainable practices. At the same time, the necessary infrastructures to accommodate the increasing needs of the developments in AI raise threats to the environment.

Environment protection

SDG data labs have been suggested as an effective method of supporting the development of SDG indicators, analysis and visualisation platforms to provide an enriched understanding of the current ecological conditions and ability to respond to environmental challenges (United Nations 2014, p.24). Beyond this, technological advancements and new forms of data collection have meant that the collection of data from non-traditional and more diffuse sources is possible (Bennett et al 2013). As understandings of behaviour develop, it is possible to gain better insights into practices which have a direct bearing on the environment. Not only can this be used to better safeguard against pollution such as energy waste captured by smart sensors, but it can also be used to develop a more robust understanding of human and environmental vulnerability by combining a vast array of data sets that would otherwise be difficult or impossible to consider manually (Lucivero 2019). Such considerations are pertinent to the SDGs such as Goal 12 on responsible consumption and production.

'Smart cities' offer an example of these benefits. By utilising data through the Internet of Things, smart cities can act to improve efficiency and performance (Allam and Dhunny 2019,

p.80) in a manner that works towards the realisation of the SDGs (particularly in relation to Goal 11 on sustainable cities and communities). Data can be sourced from across neighbourhoods and built into urban policy and management. The receipt of data from sensors can provide a more real-time awareness of behaviours, infrastructural conditions and ecological events (Allam and Dhunny 2019, p.80). Moreover, the granular level of information available to planners, will allow for an enhanced level of sustainability and resilience in urban environments against environmental threats (Barns et al. 2016). At the same time though it needs to be noted that this mass collection of real-time data raises serious concerns over intense surveillance that could be used for less optimal ends. Indeed, these potential dangers including abuse of privacy might delay and in cases even cancel plans of building such systems as was the case of the smart city project in Toronto (Cecco, 2019).

As our understanding of vulnerability becomes increasingly intersectional, it is important to assess the varied ways in which one can be susceptible to harm. Data sets such as humanitarian data can be utilised to capture this diversity and even to understand causal drivers of harm. The collection of information on gender, livelihood, ethnicity, or other demographic indicators can help to create a more three-dimensional understanding of where efforts are best directed and the nature of those responses in the first instance (United Nations 2014, p.22). Additionally, by combining environmental data with other datasets such as conflict, it is possible to identify the interconnections between domains of human insecurity, demonstrating that environmental harm is deeply tied to concerns such as the outbreak of violence and societal tensions (ICRC 2019).

Energy Consumption

Nevertheless, the collection, storage and processing of digital data has an environmental impact. Facilities central to the use of data, namely, data centres and cloud storage produce a considerable amount of pollution (Lucivero 2020). Data centres, which contain servers as well as the networking and storage equipment that cloud computing is dependent upon, demonstrate the lesser known environmental impacts of digital data storage and processing (Whitehead et al 2014, p.151). These data centres require energy to operate, through powering the facility and cooling the servers to prevent overheating. Moreover, to avoid power outages, diesel generators are also used within these data centres (Lucivero 2020). The impact of these diesel generators has been recognised for some time, with the US Environmental Protection Agency highlighting the need to reduce the impact of these generators since 2006 (Environmental Protection Agency 2007). Nevertheless, the environmental practices of digital service providers exist on a gradient of harm. The environmental protection organisation Greenpeace provides information on the sustainability practices of digital service providers, recognising that some companies operate on a 'cleaner' basis than others, seeking to mitigate their carbon footprint (Cook et al 2017, p.86).

The scale of centres together with the energy that they require to perform actual computation, is also crucial to the environmental impact they have (Luciverno 2019). In 2015, data centres were said to contribute to 2 percent of global greenhouse gas emissions (Vaughan 2015). According to more recent predictions data "is set to account for 3.2 percent of the total worldwide carbon emissions by 2025 and they could consume no less than a fifth of global electricity" (Trueman 2019). Moreover, they have the fastest growing carbon footprint across the ICT sector as a whole (Avgerinou et al 2017, p.1470). The scale of their impact, as well as their expansion and entrenchment in a data-driven society, has the potential to counteract the benefits that arise from the use of data to bolster efforts to realise the SDGs. However, in highlighting these challenges, one must not ignore the increasing momentum in developing 'green computing' and the ability of such approaches to mitigate

some of these concerns (Gai et al, 2016, p.46). Considered data deletion (see sections 2.5 and 7.4) should also be considered as an option to mitigate harms on the environment.

Distribution of Environmental Harm

Notably, data service providers have a role in the distribution of economic harms, as exemplified by Facebook relocating its servers to Iceland in order to reduce the amount and cost of energy used to cool its servers (Adalbjornsson 2019). Such a relocation is perhaps only possible for a restricted number of organisations with the ability to take such large-scale actions and operate internationally at a large scale. Cloud computing may allow smaller businesses and organisations to use the infrastructures of those who have relocated. Though it may allow for fewer data centres to be built, it may pose concerns that this will also have environmental impacts that will be concentrated in the location of the host infrastructure (that is determined by business interests) as opposed to being justly distributed throughout the globe. This ability to relocate energy intensive servers may also pose a challenge for the meaningful regulation of such activities. Indeed, it may lead to corporations shifting their facilities to areas with more favourable conditions such as cheaper energy allowing them to continue polluting (Jones 2018). Additionally, these data centres will invariably have a bearing on the local environments to which they have been relocated (Taylor 2018). This in turn raises ethical considerations, namely, how harms are being distributed globally, particularly where the decisions on their locations are being made on the basis of capitalist interest. The economic incentives offered by these organisations can often lead to conformance, and the overlooking of the environmental harms that may materialise (Lucivero 2020).

Furthermore, environmental concerns arise from the corollary equipment used for data processing. In this respect the disposal of computing hardware can be harmful when not handled in a sustainable manner. For instance, the collection of valuable materials from the hardware through copper and gold and the use of practices such as incineration, can lead to environmental pollution (Williams 2011, p.355). The 'data revolution' that is propelling the collection and use of ICT systems, whilst presenting an opportunity to better address issues such as climate change, must be considered together with the environmental impacts of the equipment used to action this effort. This is particularly the case where disposal sites are more likely to be situated in less-affluent communities (Lucivero 2020). The social justice issues that arise in relation to the distribution of environmental harms resulting from the use of digital data (and the equipment necessary for its utilisation) present furthering 'climate apartheid' whereby poorer and more marginalised communities are left to face the brunt of harmful environmental impacts, and those with better financial means are able to evade these impacts (Alston 2019, p.14).

Conclusion

Through the collection of increasing amounts of data, in increasingly real-time, our knowledge of and responses to environmental harm is improved. Necessarily, however, the environmental impacts of its storage and use must be understood and appreciated both in terms of its inherent harms as well as the balancing of the distribution of harm across demographics and locations. The ability to capture information that pertains to the environment, behaviours and vulnerabilities (including data that was not previously obtainable) can allow for better efficiency, and policy development. These developments ultimately support the realisation of the SDGs and demonstrate the possible positive outcomes of using data.

The section illustrated the considerable beneficial effects of AI in important sectors like health and government, as well as economies in general through increased efficiency and depth of

insight. Moreover, the role of AI in innovation is crucial due to the development of new products and solutions (for example within the health sector, professional, scientific or technical services). However, the underlying structural issues indicate a potential for increasing inequality, long-term economic harm and the stifling of innovation. Moreover, the source of data, data choice, inherent bias in the datasets, lack of data access, quality and characteristics of data can harm individuals, potentially marginalise vulnerable groups and amplify inequalities. The next section will discuss law and transparency as modifiers of this negative impact aim to address these issues without disrupting the growth of AI for public good.

6. Law and Transparency as Modifiers to Impact of Data in AI

AI applications have been increasingly used in every aspect of our lives and societies as discussed already. In fact, the pandemic outbreak of COVID-19 illustrated to the lay public the potential of AI in sectors like health and research whereas it supported in cases the rapid shift to digital environments in education and work. At the same time, the use of AI in these conditions exposed the potential risks emerging from the data collection, accessibility, availability, processing and sharing analysed in the previous section. Here, we discuss how law and transparency can act as modifiers to these socio-ethical, economic and environmental issues without disrupting the further development of AI.

6.1 Impact of law on data availability and data access

This section focuses on the impact that law or the lack of clear legal provisions can have on data availability and data access. Data access here does not refer only to manual access given to specific individuals but also access given to legal entities and/or algorithms. The collection of large amounts of data and the use of big datasets is necessary to train algorithms and to develop AI tools. For this reason, analysing the legal implications of the use of data for AI is important. However, access to and use of data can be limited either by privacy laws or by intellectual property rights that protect the commercial value of data for firms and individuals (Martens 2018). Similarly, the lack of legal certainty around data ownership and data sharing may hinder the unobstructed flow of data among countries and sectors. The main issue is how to find a balance between open data access that enables the development of AI tools while also respecting the rights of data holders and data subjects. These rights should be respected by people and companies using AI products, as well as by fully automated algorithmic tools operating without human intervention.

Privacy and data protection

Privacy and data protection laws require that certain safeguards that protect individuals are in place to allow the free flow and utility of data. Privacy is considered a human right at the international level (UNGA 2013), and more specific rules regulating access to personal data exist at regional and national levels (Greenleaf 2019). These rules vary among different countries and regions (DLA Piper 2020), but in general, they may limit access afforded to AI companies to datasets containing personal data. According to the GDPR, for example, the use of such datasets must comply with a set of data protection principles (i.e. processing for a specified and legitimate purpose, data minimization, consent). Likewise, privacy regulations require companies to give people access to and information about the processing of their data (Chander, Kaminski and McGeveran 2019). However, compliance with these requirements in an AI context is difficult and the effects of privacy rules on data access for AI is uncertain (ICO 2017a). The lack of explicit answers to AI-related privacy issues might lead to further

uncertainties and costs for AI companies. Increased costs may then limit their access to data and hinder the development of AI tools (Sartor and Lagioia 2020).

Despite some similarities, privacy regulations around the world follow different approaches to data access and data use due to the diverse s understanding of privacy (Simperl, O'Hara and Gomer 2020). The lack of a unified global data privacy regime can hinder the flow of data beyond jurisdictional borders. While the GDPR enables the sharing of personal data among EEA countries and with third countries that offer an adequate level of data protection, there is still tension about sharing personal data with other parts of the world where major technology companies are based (EDPB 2020). Harmonizing regional and domestic privacy frameworks via international collaboration can make data rapidly and equitably available for the development of AI-driven tools (Schwalbe and Wahl 2020). Otherwise the development of AI tools will depend on the availability of data in each separate region or country.

The differences in the regulation of data access across regions can also influence the quality and competitiveness of AI tools. AI companies may have an interest in developing their tools in countries with less restrictive privacy laws (Mercer 2020). Fewer privacy restrictions are seen as enabling AI development by increasing data availability and reducing compliance costs. Nonetheless, data protection rules can enhance consumer trust and increase demand for AI solutions developed under robust data protection frameworks. In the long term, the existence of well-respected privacy safeguards can increase the amount of data the public is willing to share for research and development of AI tools (ICO 2017b). Some states have sought to use their strong data protection frameworks to entice investment. For instance, Iceland, as discussed in the environmental section, has presented itself as a 'data haven' due to its robust data privacy laws (Gaedtke 2014).

IPR

Another legal framework with direct impact on access to data for AI are intellectual property rights, specifically copyright and database rights. These rights can both enable and hinder the creation of AI, but this evaluation is case specific. One fundamental reason is that IP rights, particularly rights in databases, are treated differently in national law and there are significant variances in the degree of protection for databases. While there are applicable international laws, these only establish minimum standards of protection that must be guaranteed in domestic law. To understand how IP regulations, affect data accessibility, we can look at it from the perspective of database rights.

A database in the U.S. can only be protected with compilation rights, which affords copyright to a database only if its data has been "selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship" (Feist Publications, Inc., v. Rural Telephone Service Co. 1991). Essentially, compilation rights only protect the databases assembled or curated with a minimum level of originality or creativity; they do not afford protection to a creator solely due to the time and/or effort that went into a database creation. In the U.S., databases that are not eligible for compilation rights are not protected by IP laws, and creators cannot lawfully prevent another from accessing or using the database under IP law. Other countries that only afford this type of protection include Australia, Brazil, China, Hong Kong, Japan, and Singapore (Wilks 2014). Consequently, it is harder to protect databases created in these countries under IP law and unprotected databases can more easily be accessed for the development of AI.

In Europe, in addition to compilation rights (EU Database Directive 2019, Art. 3), an additional form of protection exists: *sui generis* database rights (EU Database Directive, Art. 7). A *sui*

generis database right protects the creator's investment in obtaining, verifying, or presenting the contents of the database. In order to qualify for protection, that investment must be a substantial use of resources and/or effort in the qualitative or quantitative sense. In essence, this right recognizes and protects the time and resources that go into creating and/or maintaining a database. This type of broader protection enables the owner to prevent another from using the database, thus potentially limiting the ways the data may be used for the development of AI. There are lawful uses of a protected database, but it must be one identified in law (e.g. sole purpose of illustration for teaching or scientific research) (EU Database Directive 2019, Art. 6). A similar type of protection exists in India, South Africa, and South Korea (Wilks 2014). Databases created in these countries are easier to protect under IP law and may not be as easily accessed for the creation of AI.

However, it must be noted that IP law is not the only way for a database creator to protect a database. In both categories of legal regimes – with or without *sui generis* database rights – creators can lawfully restrict access to a database through contract law, for example by licensing or confidentiality provisions.

Data ownership

Beyond IP and privacy rights, the availability of data for AI raises questions around the concept of data ownership. The legal uncertainty around this concept has been recognized as a potential barrier to the use and free flow of data (EC 2016). Data ownership is used by AI and big data stakeholders to claim some exclusivity over data and/or databases, when other rights (e.g., IP law) cannot be applied (Van Asbroeck et al. 2019). Such data ownership claims are normally made to ensure that an individual or an organisation can use their data and can take advantage of the benefits from using them. However, as ownership relates mostly to an understanding of data as property, there are significant problems that make it less suitable for use in the AI context.

Digital data are not like tangible things that can be used by only one person. Data are non-rivalrous and non-excludable, which means that they can be duplicated and used by lots of people at the same time. Data can also be in cases non-depletable as they can be used many times without losses in quality and value (Hummel et al 2020). In addition to these issues, the concept of ownership is problematic particularly with regards to personal data, since the latter retains a link with the data subject even if it is somehow transferred to another entity. Data can also contain information about more than one individual (e.g. genetic data). In such cases it is not clear who should be the rightful owner of the (personal) data in question (Leyser and Richardson 2018). The prevailing view, therefore, is that the idea of data ownership lacks a well-established legal basis in many jurisdictions (Hummel et al 2020). Similarly, data ownership is not considered the best way to protect privacy and personal data, as it gives individuals the option to trade away their privacy rights and reduces these rights to commodities.

Since ownership rights are not suitable for regulating access and use of data, the concept of rights in co-generated (personal and non-personal) data has been developed as equivalent to data ownership (ALI and ELI 2018). This concept includes not only rights deriving from a data protection framework (e.g. to access or transfer the co-generated data, and to have them corrected), but also the right to enjoy an economic share in profits derived from those data. To be entitled to these rights, different ways of participation in the generation of the data have been suggested, such as being the subject of the information or the owner of the object of the information (ALI and ELI 2018). It is important to note that, contrary to the concept of

data ownership, the aim behind the development of these rights is to make them functional and not exclusive, in order to allow for the broad sharing and use of data (Ducuing 2020).

Data sharing

The availability and flow of data for innovative use and for the development of AI depends also on establishing a legal regime governing the relationship among different data holders or controllers – the physical or legal entities that can access and use specific datasets. A prerequisite to data sharing is also the creation of datasets according to a set of agreed standards that will contribute to the quality and availability of data (Reimbasch-Kounatze 2015). Regarding the sharing of data among different public or private actors, on the one hand, emphasis has been given to making public sector data available for use by private actors. In the EU and in the UK, for example, relevant legislation ensures that the public sector makes most of the data it produces easily accessible for use, not only by private companies but also by civil society organizations, and scientific researchers (EU Open Data Directive 2019; UK Open Government License). On the other hand, the importance of enabling private companies to share the data they hold with public authorities has been highlighted. According to the EU's data strategy, there is currently not enough private sector data available for use by the public sector. By accessing such data, public authorities can develop or use better AI tools to improve policy making and the general quality of public services (EC 2020a).

In addition to standardizing data flows between public and private actors, building legal frameworks that shape the sharing of data between private entities (business to business) is equally important (EC 2020b). So far, the lack of legal clarity on who can do what with data (e.g., with data in mixed datasets containing private and non-private data) has slowed down data sharing between businesses. For example, resolving issues related to rights in co-generated data, such as data for AI in industrial settings, and clarifying the legal rules for a responsible use of this data can support business-to-business data sharing (EC 2020a). Finally, fragmentation between public authorities at a national and a regional level should be avoided. Data sharing between public authorities can make a considerable contribution to improving policy making and public services. In the UK, the Digital Act of 2017 provides the legal basis for government departments to make use of digital data and allows data sharing among government departments.

Conclusion

The existence or lack of different legal frameworks governing access and rights to data and datasets can significantly impact their availability for the development and use of AI tools. Stricter privacy and IP rules can limit the use of data, but they can also create incentives for individuals and organizations to share their data or to engage in building new tools and/or datasets. The lack of harmonized global privacy or IP legal regimes, however, may hinder the cross-border flow of data and, thus, restrict the development of AI at a regional or national level. At the same time, the legal uncertainty around data ownership prevents data holders from engaging fully with their data. Providing clarity to the rights of data holders and establishing concrete rules governing data sharing among public and private actors can increase the flow and availability of data for AI.

6.2 Transparency for data and AI

In the context of data, transparency means that data can be accessed, processed, understood, deleted and presented easily. It needs to be acknowledged that what can be transparent for one individual is not necessarily transparent for another. Moreover, transparency is not a good or ethical value in and of itself, rather it is a means to support other values and benefits, such as autonomy, responsibility, and accountability (PFST 2019). Transparency can support

users in better understanding the potential beneficial and harmful implications of the use of their data, insights derived by, and decisions made by systems that use that data, including AI. This is both at an individual and community level. It can also make it possible to identify and address when laws have been breached. This can increase their opportunities for respectful, dignified, and trusted interactions with those the data and the AI is intended to serve. Whilst transparency is traditionally connected to governance, the role of the private sector when it comes to data needs to be underlined. For industry, transparency is essential to its efforts to persuade government and society that it can commendably self-regulate, and to show that they do not perform actions that go against the public sentiment, morals or of course, privacy and data protection laws. In turn, for government and its data protection authority branches, data transparency is required to facilitate proactive monitoring activities, as well as any investigative activities

Acts of transparency range from independent audits of the datasets to new regulatory frameworks that support supervision. However, ultimately transparency also depends on the user's capability. 'As such, transparency should always address the interplay between those who provide open data, the functionalities of the system that enable access to the data and the open data user' (Zuiderwijk et al. 2014).

Transparency of what?
One of the biggest challenges for transparency is identifying what needs to be transparent regarding:

- The nature of the data itself
- The source of data, including who collected it and how
- What the data show
- What happens with the data?
- How the use of that data is decided and what governs the decision-making process

To answer these questions, data needs to be readily accessible, and we also need to understand who was involved in determining the requirements around the data and its possible use in an AI application or elsewhere. Specifically, when data is used for AI, its use should also include rendering relational aspects between the data and the AI more accessible:

- What logic and models drive the relationships between the data and the code?
- Which data are highlighted as decision making variables?
- How do the outcomes relate to the question being asked of the data?

This process is more than just record keeping but it is also about ensuring the appropriate audience has clear information about an AI system's capabilities and limitations, including the data that has been used to train it (European Commission, 2020).

Transparency begins with the data
Transparency efforts should include data quality assessments prior to initiating any data driven project, including AI, to determine whether the target of the project or AI match with the data being used. Assessments should include the following considerations: what data was used; why that data was used; how the data was deemed appropriate and acceptable evidence for answering the question the AI is addressing. It includes being up front about the raw data and sources of the data, how data is pre-processed, how it is verified, how updated the datasets are, and how an AI is retrained if data changes (PFST 2019). This kind of

assessment is especially important as AI increasingly uses Neural Networks, which are black boxes that sometimes make understanding how or why a decision got made nearly impossible (PFST 2020).

Focusing on these initial questions – the human context behind the data and algorithms – in terms of operational goals, inputs, outputs, and especially outcomes is an important form of transparency (Kroll 2018). This involves being able to explain the assumptions made when gathering the data, as well as how the data was deemed to reflect the world and the object of the AI sufficiently and validly.

But it also needs to step away from the data to ask "for whom" is the transparency and "why" is it needed?

This is because transparency has many valences. Some important ones are:

Explicability.
This refers to the interpretability of an AI system by users and the public, so they can understand how decisions are made by AI, what is considered and why. It is the ability of a user to know how an algorithm works and why it produced the outputs it did, in a specific context (PFST 2019). As importantly, explicability refers to users' or publics' ability to understand how the rationale behind the AI relates to and influences their practices and decisions with the algorithmic outputs. Thus, explicability needs to also empower users by pairing what is offered as explanation with appropriate training and digital literacy skills (Council of Europe 2019). This is especially important for the public. While chains in data relations could be explained to the public, they would take resources and skills most people don't have (Oswald et al. 2018).

In addition, "explanations should be *socially meaningful. T*he terms and logic of the explanation should not reproduce formal characteristics of the models/analytics or technical meanings and rationale for the models. They should be "understandable in terms of the societal factors and relationships that the decision or behaviour implicates" (Leslie 2019, p. 36). This is increasingly prescient, as the more in depth an explanation is on technical matters and statistics used, the more data science background someone needs to understand the explanation.

Justifiability.
This refers to gaining enough of an understanding of a tool for a user to justify a) how the tool was designed, b) how the AI is implemented as part of a broader decision-making process, and c) how the outcomes are used to make specific decisions (AIHLEG 2019). This should include how the AI approaches ethical issues (e.g. how does it approach fairness, how does it support checks for bias, how did the designers and users continually ensure the AI does not discriminate), in context, and why a user or the public should trust the outcomes of an AI. If the rationale behind AI is not articulated, meaningful consent to data processing is not possible, and any challenge to a decision becomes difficult to legitimise (Mittelstadt et al 2016).

Accountability.
This refers to the ability to understand how an algorithm works,the ability to explain different properties of the AI system, product or service and how it can be used. Opaque AI, and thus, opaque use of data, make it difficult to identify if and where laws have been breached, including human rights laws, and how to attribute liability or responsibility (European

Commission 2020). The transparency mechanisms put in place need to be carefully balanced with the context of use and what regulations exist (or need to exist) to support liability actions.

Traceability.
This refers to the ability to backtrack on a particular function or insight in order to understand how it was derived and the logic through which it was produced. It includes being clear about which datasets were used by the AI for which decisions, as well as how they were gathered, labelled, and cleaned prior to training the AI (AIHLEG 2019). This is important for attributing responsibility, and "ensures that a system's operation can be explained from the tracks it leaves, hence the quality of explainability" (Cerna 2018, p.18). This is important not just for justifying actions taken, but as implications for broader data practices as well. For example, without the ability to trace back how a decision was made using the data, it is difficult to know where the responsibility lies for a wrong decision. Without that, it is nearly impossible to identify a correct pathway for rectification.

Making sure Transparency helps not harms
It is also important to note that transparency can sometimes come with costs or produce harm. This harm will need to be measured according to the principle of proportionality, in other words, one must assess whether the harm outweighs the good. It is not always clear to whom information should be made transparent (Floridi and Taddeo 2018) as sharing information for the sake of transparency could do more harm than good. For example, if the data contains personal or sensitive information, such as medical data, making it transparent to the public could cause greater harm than the accountability it provides. Similarly, while making data and processes transparent allows for the public's assessment of how the data was used it also has the potential to increase certain groups' vulnerability, depending on what and to whom is revealed (AIHLEG 2019).
These also include the added staff training and resources necessary to ensure the explanations are meaningful and fit-for-purpose. Similarly, regulatory bodies need to be in place to oversee compliance. Transparency also has to be balanced with trade secrets, and businesses need actionable oversight protocols. Even more, "public access to data can flame interest-driven controversies via the untutored or unscrupulous misuse of data", make algorithms more readily hackable, or can lead to unintentional privacy breaches (PFST 2019).

Conclusion
The section discussed the complexity, challenges and risks of global legal frameworks related to data governance. It concluded that establishing concrete rules to govern data access and sharing among public and private actors can benefit the development of AI and increase the flow and availability of data. The responses to these issues must be properly addressed and appreciated through policy and practice. Ultimately, using data and AI should adhere to the principles of: respect for human autonomy, harm prevention, fairness and transparency to maximise positive outcomes and minimise negative socio-ethical impacts.[8] Transparency is a concept that can lead to questions of current practices, provide opportunities to address the aforementioned issues and lead to policy recommendations and best practices.

[8] European Commission, High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (2019), p 12.

7. Availability of and accessibility to data for AI development: data quality and challenges in three fields.

´Governments should also consider public investment and encourage private investment in open datasets that are representative and respect privacy and data protection to support an environment for AI research and development that is free of inappropriate bias and to improve interoperability and use of standards.' (OECD.AI, 2019:2.1(b))

Availability and accessibility of data for AI development is of key importance to realise the full benefit of technology development for communities across the globe. However, openness is only one aspect of accessibility. Accessibility also refers to discoverable, good quality, timely, and fit for purpose data. Furthermore, data management within AI projects and sound governance at organisational levels are of key importance so this can be realised and more data can be made openly accessible.

In this section the key characteristics of quality data, discussed in Section 4, will be further elaborated drawing specifically on three fields: AI used for rapid diagnosis and drug discovery in Pandemic response, AI for developing Human Language Technologies and the use of AI in the justice system. For each example, we provide an overview of current work in this area, the characteristics of data needed for optimum results, current challenges related to data use, and key initiatives on addressing these challenges. The section concludes with recommended actions for each stage of an AI project with respect to data, so that more quality data is available for the development of AI applications.

7.1 Development of AI in the Pandemic Response

Case Study: Accessing timely data of high quality

Pandemics require a collaborative and rapid response in the form of diagnostics, tracking the spread and developing treatments (including drugs and vaccines). In the case of COVID-19 we have seen that AI plays a very important role in this process due to its ability to analyse vast amounts of data to detect patterns and develop predictions. AI technologies have been used, for example, in developing rapid diagnostics as well as medication, through drug repurposing.

Developing AI for rapid diagnostics requires vast amounts of multitudinal and multimodal data for training purposes; the data must also be clean and annotated so that classifiers can be well trained. Limited data will skew the result and allow for less accuracy in terms of diagnostics. In the response to COVID-19, AI systems were developed, which can detect coronavirus infection by analysing CT scans or X-rays of the lungs. (Santosh, 2020). These tests are very important as access to either X-ray machines and CT-scans are widely available in health care systems around the world, especially while the RT-PCR tests were in short supply. The system could identify a potential COVID-19 infection and thus prioritise analysis for the physicians and reduce the overload. However, the refinement of the system for providing an accurate diagnosis was hampered by lack of data, especially data annotated by physicians (Ray, 2020).

Drug development is a lengthy and resource heavy process which includes long trial phases. To respond to a pandemic, a shorter process is needed, but one that produces results that are safe for use. Repurposing chemical compounds or drugs that have already gone through trials and are on the market is therefore a good alternative and has proved successful in the response to COVID-19. AI is optimal for this purpose due to its ability to analyse vast amounts of drug data to find chemical compounds or already available drugs.

This case study demonstrates the importance of data accessibility, data quality and that data is accessible and FAIR so that a rapid response can be undertaken when a pandemic occurs. Good data governance can help to ensure that data sources are reusable (through use of standards, good practices and compliance with the FAIR principles), of good quality (with metadata containing information about AI-specific characteristics of the dataset), and as open and accessible as possible. There is a clear role for collaboration across sectors for ensuring that the data each sector holds (e.g., health data, socio-economic data, drug and chemical compound data) is of good quality and that it can be accessed with minimum delay.

Health data for developing rapid diagnostic AI in a pandemic

AI has the potential to assist rapid diagnosis in pandemic cases, which is key to limiting contagion and understanding disease spread. To develop any model and algorithm, vast amounts of varied training data are required in the first stage of the process. To this end, it is crucial that such data is updated given the pandemic progress, so the response is appropriate to the real-life situations. Furthermore, the data must be representative of the population to ensure scalability and accuracy and minimise risks of bias. Clinical needs change at different stages of the pandemic and data and models must reflect this dynamic process. In the case of COVID-19 development of AI tools has assisted with rapid diagnosis using image and symptom data (Santosh, 2020).

Characteristics of the datasets

- Timely datasets are required to allow for a rapid response. Accessibility, openness and FAIRness are key in realising this characteristic.
- Datasets composed of varied and multimodal data (e.g., imaging, symptom data, health records, drug interventions, location data, etc.) are required as this increases prediction accuracy.
- Well-described and accurately labelled data is needed for supervised and semi-supervised learning.
- Unlabelled data can be used for unsupervised learning, but this requires human oversight and an understanding of potential limitations of the outcomes.
- Data from different health systems and institutions needs to be integrated to create large datasets for training and evaluation. To allow for integration and interoperability, standardisation of formats, ontologies, vocabularies and metadata are of key importance.
- Datasets must be up to date as they require constant time-specific data to reflect different stages and developments of pandemics. If datasets are inactive, this information should be clearly made available in the metadata to assist researchers with avoiding using out of date data.
- Datasets must be representative of the whole population to ensure scalability and accuracy as well as to minimise risks of bias. This can be difficult to achieve, especially at the start of a pandemic, but data should be continuously updated, and care should be taken to incorporate data from different groups and locations.

Current Challenges

- Data sharing efforts in health are fragmented (Luengo-Oroz, et.al., 2020) and can be slow due to institutional and legal barriers. Health data is personal data and often classified as sensitive or special categories of personal data (e.g., under GDPR). This places certain demands on data custodians, which render openness and sharing of data a complex, costly and slow undertaking.
- There has been a lack of multitudinal and multimodal training data on coronaviruses, which has led to AI tools presenting skewed results (Santosh, 2020).
- Some health data is not sufficiently FAIR (OECD, 2020c) and in formulating a response to COVID-19 researchers have found that 'there is a confusing plethora of publicly available COVID-19 surveillance data resources. Relevant websites are frequently poorly designed making it extraordinarily time-consuming and frustrating to find and extract the relevant information.' (Austin et al., 2020:1)
- Pandemic data is uneven in terms of coverage. Countries with under-resourced healthcare systems may lack staffing and funds to ensure data governance practices at every stage of the data lifecycle – this includes structured collection, data curation work, safe storage, etc. This means that data from some countries is missing from datasets about pandemics (Cornish et al., 2020).
- There is a high risk of bias in health data, as only people who come into contact with the healthcare system are included and people who do not seek medical attention are missing. In many instances this accounts for people on low income and people who lack access to healthcare.

Examples of work on data accessibility in this area

COVID-19 Open Research Dataset Challenge (CORD-19). This dataset is a 'resource of over 200,000 research articles, including over 100,000 with full text about COVID-19, SARS-CoV2 and related coronaviruses. The dataset is provided to the global research community https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Google Health has developed a COVID-19 Open Data repository, which is 'a comprehensive, open-source resource of COVID-19 epidemiological data and related variables like economic indicators or population statistics from over 50 countries. Each data source contains information on its origin, and how it's processed so that researchers can confirm its validity and reliability.' https://github.com/GoogleCloudPlatform/covid-19-open-data

Virus Outbreak Data Network (VODAN) A joint activity between CODATA, RDA, WDS and GO FAIR, which proposes to create FAIR data repositories for machine readable, interoperable and reusable clinical data, which can be used by incoming algorithms to ask specific research questions. This allows for rapid access to data while also respecting privacy. *https://www.go-fair.org/implementation-networks/overview/vodan/*

RDA Working Group on Sharing COVID Data. This expert WG was set up to provide recommendations and guidelines for rapid and secure data sharing to meet the needs for a coordinated global response to COVID-19.' https://www.rd-alliance.org/group/rda-covid-19-epidemiology-rda-covid19/outcomes/sharing-covid-19-epidemiology-data

Chemical and drug data for AI assisted drug discovery/drug repurposing

'In the big data era, artificial intelligence (AI) and network medicine offer cutting-edge application of information science to defining disease, medicine, therapeutics, and identifying targets with the least error.' (Zhou, 2020: 1) In this context, drug repurposing has become a promising approach in medicine as it brings an opportunity to reduce development time and overall costs considerably. AI has already been used in the pandemic response to COVID-19 analysing drug data and previous findings. The results indicate that existing drugs, such as Remdesivir, which was initially developed as a potential treatment for Ebola has shown great promise in the treatment of COVID-19 (Ibid:7).

Characteristics of the datasets

- Data will need to be from disparate sources such as real-life data (electronic health records), chemical compound data, cellular data, trial data, etc., but interoperable and harmonised into unified databases to guarantee a broad applicability across different scenarios.
- There is a fundamental need for high quality, large and clean datasets so that AI can be successfully used to identify chemical compounds and potential drug combinations that can be used for repurposing drugs for COVID-19 (Wakefield, 2020). For personalised drug repurposing (which greatly improves disease treatment) massive genetic and genomic data are required, in addition to the data listed above.

Current challenges

- Too much data is siloed within individual companies, e.g., large pharmaceutical companies or within labs at research universities. Bringing this data together is hindered by legal issues (e.g., IPR), commercial interests and administrative barriers (Ibid).

- Real world data, such as electronic health records are often of lower quality (e.g., is often incomplete) and have higher dimensionality (including confounding factors). (Zhou, 2020:5)
- Data heterogeneity and low quality are presently a barrier and slow down progress in the field of drug discovery.
- Lack of standardisation and harmonisation of data to form a unified database to allow for machine learning approaches.

Examples of work on data accessibility this area

American Chemical Society division, CAS, has released a dataset containing 50,000 compounds with potential antiviral properties to support the discovery of drug treatments for COVID-19. The dataset is open source and the license terms support use for applications including research, data mining, machine learning and analytics. https://www.cas.org/covid-19-antiviral-compounds-dataset

NIH NCATS Pharmaceutical Collection is a publicly accessible collection of approved molecular entities which provides a valuable resource for both validating new models of disease and better understanding the molecular basis of diseases and interventions. It consists of nearly 3,000 small molecular entities that have been approved for clinical use by U.S., European Union, Japanese, Australian and Canadian authorities. https://ncats.nih.gov/expertise/preclinical/npc

OpenTrials is a collaboration between Open Knowledge International and Dr Ben Goldacre from the University of Oxford DataLab. OT aims to locate, match, and share all publicly accessible data and documents, on all trials conducted, on all medicines and other treatments, globally. https://opentrials.net/

7.2 Human Language Technologies for under-resourced languages

Case Study: Lack of available data for development of AI technologies

Human Language Technology (HLT) refers to the production of technologies that seek to understand and reproduce human language. All these technologies produce tools that are used in a range of fields, e.g., communication, health and education, etc., and can significantly improve people's quality of life.

The development of HLTs takes a vast amount of language training data; however, the amount of data available for languages is extremely uneven. The term under-resourced language refers to a language that displays some of the following characteristics: lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, lack of electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, and vocabulary lists (Krauwer, 2003; Berment, 2004).

HLT requires considerable amounts of corpora of text, audio recordings (including transcriptions), and dictionaries (some of which are manually annotated) (Crystal, 2000). For example, to produce a natural sounding voice through speech synthesis (text-to-speech), at least 30 hours of recorded speech are required. Speech recognition systems (speech-to-text) are even 'hungrier' for data, requiring at least 300 hours of recorded speech (some systems use over 2,000 hours) and large vocabularies of over 60,000 words. Developing data sources for under-resourced languages is thus very resource intensive, which is a clear barrier to development of HLTs for many languages.

This case study is an example of how the cost and effort of creating data for AI may impede the development of technologies for specific language groups, limiting their access to services in their own language. Data governance can assist with ensuring that language data that exists is curated with re-use in mind and made both FAIR and open as far as it is possible. With respect to data creation, there is a clear role for collaboration between sectors (public, private and academic) in creating language resources and making them available for AI development.

Characteristics of the datasets

- Large amounts of different types of data are required to develop a HLT in a specific language, such as audio files and text data, including phonetic alphabets and lexicons.
- Audio data needs to be of good quality in the sense that voice and pronunciation are clear and background noise is kept to a minimum. Having trained voice talent for recording is optimal at this stage.
- Data needs to be structured, clean, labelled and free of errors.

Current challenges

- The cost of creating a data corpus for a language, where data is not readily available is very high and requires extensive manual work and human input.
- If lexicons are to be created, these need to be manually checked. The creation of lexicons and language data is extremely time consuming and costly.

- In many instances audio data that is scraped from the web is of low quality and needs considerable preprocessing before it is usable for processing and analysis.
- With regard to scraping audio, video and text data from the web, the ownership of the data and privacy of individuals who produce them pose complex legal and ethical questions.
- Many languages have a variety of regional accents and dialects, in addition to the multiple accents originating from second-language speakers. Newer AI technologies are working to solve this issue to some extent by 'triangulating the phonemes' to provide more specific results (Stoltzfus, n.d.). Consideration for accents and dialects should however still remain a consideration at the data creation/collection stage.
- Data for under-resourced languages is not available to the same extent as the more dominant languages, such as English. This poses a significant barrier to the development of HLTs for these language groups.

Examples of work on data accessibility and availability this area

CLARIN - European Research Infrastructure for Language Resources and Technology has the mission to create and maintain an infrastructure to support the sharing, use and sustainability of language data. CLARIN currently offers 12 types of corpora, lexical resources and language data tools. https://www.clarin.eu/

Mozilla Common Voice is part of Mozilla's initiative to make more voice data available for AI development. Anyone can submit a recording of their voice and through this crowdsourcing initiative thousands of hours of speech for 60 different languages have been collected. https://commonvoice.mozilla.org/en/about

7.3 Data for developing AI applications in the Criminal Justice System

Case study: Different lifecycles of public sector data and challenges for AI development in the justice system- an example from the UK

The use of AI in the justice system promises more consistency of court decisions, objectivity, increased efficiency and quality of justice (European Commission for the Efficiency of Justice, 2018). However, decisions made on the basis of AI and data analytics will have direct impact on people and communities and therefore there is more emphasis on sourcing data responsibility, being aware of bias in data and therefore in the delivery of justice (e.g., in sentencing and conflict resolution). Much of the data used in the justice system is personal and sensitive data. Therefore, there is increased need for transparency, impartiality and equity in the process, as well as human oversight and independent expert assessments to validate decisions (Ibid)

While not only focused on the criminal justice system, a recent £1M GBP investment was made by the United Kingdom government into the digital transformation of the country's Courts and Tribunals Judiciary. Key challenges related to this work were linked to the different 'lifecycles' of data processing within AI-powered justice processes (Aidinlis et al.,2020)

More specifically, the legal and ethical implications of using data in AI-driven justice can be structured around four distinct, yet overlapping, stages of data processing: (1) collection, (2) preparation and linkage, (3) access and (4) retention/re-use. These stages cover the whole spectrum of data processing within a data infrastructure that will be used as the key resource for developing sophisticated algorithms in the justice system. Before automation can happen, public bodies will have to collect significant amounts of data, curate and clean them so that they can be fed into algorithmic training systems, link them with datasets belonging to other public bodies and think strategically about the future retention and re-use of data.

This case study demonstrates that there are noteworthy legal and ethical implications related to data 'lifecycles'. When collecting justice data, authorities should be taking all reasonable steps to both maximise potential benefits and mitigate relevant risks. When preparing the datasets for linking with other datasets and use within algorithms, public bodies should ensure that appropriate de-identification measures are applied so that personal data of citizens is accessible on a strict need-to-know basis by certified personnel. Access to justice data should be governed by reference to the contribution of a particular use of data to the 'public interest', and under conditions of monitoring by dedicated governance committees.

Characteristics of the datasets

- Data especially for judicial decisions should come from a certified source and no modifications should be made until after the learning stage. If any modifications are made, these must be documented carefully in order to allow for full traceability and transparency in order to maintain trust in the system, and versioning must be clear.
- Principles of transparency, impartiality and fairness need to apply to the selection, the quality and organisation of the data (Ibid: 11).
- Datasets must be diverse and need to reflect societies and communities which AI systems affect. (House of Lords Select Committee on Artificial Intelligence, 2018: 43)

- Data should be accessible to professionals in the justice system for review.

Current Challenges

- Data on criminal arrests and convictions can be biased as they originate from biased justice systems. Studies from the US have demonstrated that Black and Latino people are more likely to be arrested and convicted than white people due to systematic bias in law enforcement (Stevenson and Mason, 2018; Mitchell and Caudy, 2013).
- Data on arrests has been found to be "dirty" due to its links with performance measurements for different police districts in the US. In particular, a study found that systematic data manipulation of crime statistics had occurred across multiple jurisdictions (Richardson et al., 2020).
- Creating large datasets for training from disparate sources (e.g., socio-economic data, housing data, educational data) has proved to be a challenge due to differences in data collection methods and documentation across different municipalities.
- Much of administrative data has been collected without a specific purpose and while it is good for administrative use, its quality and fitness for use must be assessed before using it for AI applications in the justice system.

Examples of work on data governance in this area

The Human Rights, Big Data and Technology Project - The project is based at Essex University's Human Rights Centre with partners worldwide. It considers the challenges and opportunities presented by AI, big data and associated technology from a human rights perspective. One of the research streams is dedicated specifically to law enforcement analysing the implications of police data and technology on human rights. https://www.hrbdt.ac.uk/law-enforcement/

Data Justice Lab - The Data Justice Lab is a space for research and collaboration at Cardiff University's School of Journalism, Media and Culture (JOMEC). It seeks to advance a research agenda that examines the intricate relationship between datafication and social justice, highlighting the politics and impacts of data-driven processes and big data. Specifically, one of their projects was on Data Scores as Governance: Investigating uses of citizen scoring (funded by the Open Society Foundations). https://datajusticelab.org/

7.4 Data management for supporting AI development across different fields

The discussion regarding the data in AI development in the three different areas presented above shows that many challenges are common across different fields. In cases where data is available, it may be commercial and thus not accessible. Furthermore, available and accessible data may be of low quality, or need harmonisation to be useful in tackling complex challenges. Available and accessible data may also not be discoverable due to lack of metadata and the context of collection may be unclear due to lack of provenance. In many instances, these are issues that arise outside of the AI development cycle, especially when projects rely on data created by public and private organisations.

Employing robust data management in AI projects can mitigate many of the challenges listed above and can help to ensure that data created/collected by AI projects is of good quality and fit for reuse. For this purpose, in this section, we provide guidelines for good data management practices for AI projects, that can also assist to assess data from disparate sources before use. These recommendations can also serve to inform data governance policies within organisations.

An overarching recommendation to those who oversee and manage AI projects is to create a Data Management Plan (DMP)[9] or similar documentation for each project and we will outline the benefits of this practice throughout the AI development/data lifecycle. DMPs are a tool that can assist data creators/collectors within projects to think about the data involved at each step and how to ensure it is managed to best standards, whether the intention is to publish it for open access or to re-use within organisations.

Recommendations on data management within AI projects

Data Creation/Collection

- Consider re-use and/or data sharing from the start of the project. A key question here is what information do secondary users of the data need to be able to process it for other projects/purposes?
- Use a DMP or similar documentation to plan for each step of the AI development process and use it as a checklist to ensure good data management processes from the start. A DMP or its equivalent should be a living document that is updated throughout a project.
- If you intend to share the data openly after the project ends, consider using a data repository. Assess different repositories at this stage to ensure that you are aware of any specific requirements and standards they have for data deposits and use these to guide your data work, e.g., with respect to metadata standards, licencing and provenance.
- Assess all collected data with regard to its fitness for use and quality and document any changes made for sake of having a clear provenance. Also, ensure that you are aware of any ethical or legal issues related to your data and ensure appropriate safeguards are in place.
- Ensure that all data you collect is of good quality and record the steps and actions taken throughout the creation/collection stage. This is important for the creation of metadata and provenance.

Data organisation/refinement, processing and evaluation stages

- The data work occurs within these stages and any errors or previously un-detected data quality issues may reveal themselves here. It is important that these are attended to (e.g, for the correction for any bias found in the data). Here, you may also need to change data labels or supplement the dataset with additional reference data.
- Use timestamps and versioning methods to ensure that different versions of the data are clearly labelled.
- Use the DMP to record any actions taken at these steps that impact on the data, such as any modifications, use of different versions, corrections, etc.

[9] Data Management Plans (DMPs) are often used in research projects to document and plan for the use and preservation of data in any one project. The DMP typically describes what data will be collected/created in the project (volume, type, content, quality and format of the final dataset). It outlines the metadata, documentation or other supporting material that should accompany the data for it to be interpreted correctly. The DMP lists what standards and methodologies will be utilised for data collection and management. The plan also states the relationship to other data available, e.g.existing data sources that will be used by the project, gaps between available data that are required for the project and the added value that new data would provide in relation to existing data. DMPs furthermore note any legal and ethical issues related to any of the data sources used. See further information at: https://www.dcc.ac.uk/guidance/how-guides/develop-data-plan#Why%20develop

Curation/Preservation

- Make sure that your data has all the documentation and metadata, and is FAIR, so that other users can find and re-use your data.
- Consider the use of Persistent Identifiers for you and your data so that it is clear who created the data and so that data can be cited and you can be contacted by secondary users in case of any questions.
- Consider the use of a licence for your data so that users are aware of what use is permitted and which is not.
- If you are storing your data in a repository, make sure you have all documentation they require at the time of deposit.
- If you are storing your data on-site, make sure it is secure and appropriate for the data you are preserving.

Deletion

- For this decision to be made, a considered appraisal must be undertaken before considering deletion (Whyte, 2014).
- Do not just decide to keep everything, "just in case". Consider the space and resources needed to securely store your data and the work involved in curating it.
- There may be legal reasons for deleting/retaining specific data that you have created. Make sure you seek advice on this. These may regard specifically personal data protection, and legal retention periods.
- Consider the value of the data with respect to, number of copies held elsewhere, historical relevance, potential for re-use etc.[10]

While the data management recommendations above apply mostly to the project level, organisations who oversee data and AI development should use these to develop data governance frameworks, necessary infrastructure and guidance for its employees and contractors. This includes appropriate storage, guidance on persistent identifiers, FAIR data and metadata standards and support on legal and ethical issues. Data retention guidance and schedule will also be needed to advise developers on how to securely conduct data deletion.

7.5 Unavailable data, legal and commercial challenges

The recommendations above apply to AI projects who work with data and they assume that data exists and can therefore be managed. Different work is needed to plug data gaps, which exist due to lack of resources (e.g., language data for under-resourced languages) and expertise in preparing and integrating data and making it ready for use. As it is evident in the pandemic case, health data exists, but for it to be useful for addressing a complex, distributed health challenge like COVID-19, it works best if it is integrated with other data. Data integration is a complex task, especially in light of lack of standardisation of data with regard to formatting, as well as vocabularies and ontologies used. Integrating disparate sources of data together where these challenges exist requires extensive time and resources. There is a clear role for governments, charitable organisations as well as the private sector who support research and innovation, to provide funds and expertise in solving the challenges of data interoperability and availability.

Conclusion

This has presented three case studies which illustrate challenges regarding data availability, accessibility, interoperability and quality. Although the cases are drawn from different fields,

[10] These recommendations are based on recommendations made by the Digital Curation Centre in a number of guidance documents, which are made available on the DCC Website: https://www.dcc.ac.uk/guidance/how-guides

it is evident that many data issues are cross cutting and manifest within different stages of the AI development process. Effective data governance throughout the data lifecycle, from creation/collection through to preservation, will assist in mitigating and overcoming these challenges. The Data Governance WG and GPAI more broadly can take a leading role in developing and disseminating best practices in this respect. Section 8 will provide targeted recommendations on priority actions which can assist the WG in organising their work going forward.

8. Recommendations to the Data Governance WG to further work on data governance for AI

As stated in the foreword to this report, the mandate of the Data Governance Working Group is to *"collate evidence, shape research, undertake applied AI projects and provide expertise on data governance, to promote data for AI being collected, used, shared, archived and deleted in ways that are consistent with human rights, inclusion, diversity, innovation, economic growth, and societal benefit, while seeking to address the UN Sustainable Development Goals."*

This report, as part of this work, was commissioned to provide a description of the role of data in AI and highlight harms that arise from sub-optimal data practices as well as insufficient access to data. The report has provided an insight into key challenges in this respect and highlighted good data practices and initiatives that work to overcome these. This section will provide recommendations for action, on the basis of the findings in the report, which are meant to guide and inform the next phase for the WG as they 'identify programmes and projects that align with GPAI's mission, and could be funded by GPAI's members and in partnership with others'. The recommendations are written as concrete suggestions to assist the WG with the collation of evidence, provision of expertise, and for selecting applied projects for support so that work-specific recommendations can be undertaken.

As a general note to guide the Data Governance Working group's next steps, ongoing work on data governance as well as enhancing data accessibility within the field of research data and open government data along with associated initiatives[11] should be reviewed to minimise duplication of effort. This review will allow identifying gaps in the body of existing work and further direct the group's future efforts. In this context, we also recommend that GPAI survey their expert membership on the prevalence and applicability of data governance issues, as well as challenges with regard to data availability and accessibility, as this will allow for prioritisation of work/domains in realising data governance for AI development. We note specifically the work of the Pandemic WG, which has submitted recommendations, some of which are around governance of data to best support AI facilitated response to pandemics, such as COVID-19. Work between the two groups could in the first instance focus on this field, and then be scaled up to apply to other AI application domains.

Recommendations to the Data Governance WG

In line with their remit, we recommend that the Data Governance WG continue to work to shape best practices and standards for Data Governance in AI, through targeted research and the writing of guidelines for data use in AI projects and systems. For the development of tools and mechanisms, to further support AI developers in using the guidelines, we recommend the WG collaborate with initiatives working on the topics outlined in the recommendations.

Recommendation 1: The Data Governance WG should work to shape best practices and standards for data governance with the aim to drive access to good quality data for AI projects and systems. Actionable steps include:

Action 1a: Create guidelines around **data management** for AI projects and systems, which take all steps of the AI development process into account, from data creation and collection

[11] A list of selected international initiatives working on challenges identified in this report can be found in Annex A.

through to preservation and deletion. The WG should also work towards creating a data management plan template for AI projects and systems, which will allow for the capturing of information necessary for supporting discoverability, documentation, characterisation, trust and transparency (see recommendations 1b-1e), all of which will drive enhanced and informed re-use of data for AI.

Action 1b: Support good practices around deposition and cataloguing of AI data sources so that they are better **discoverable and accessible.** This work should include a focus on:

- Conducting a feasibility study around different options for enhancing data access for AI projects and systems. Options may include e.g., setting up a specific AI data repository or a metadata catalogue, or creating a network of existing repositories and a single discovery and access point.
- Working with initiatives that are driving the adoption of the FAIR principles, as well as the Open Science movement, and ensure that AI has input on any issues that are specifically relevant to specific data practices within the field.
- Working with the Pandemic WG on implementing their recommendation for a Central Pandemic Response Portal.[12] Lessons learned from this collaboration can then be carried forward and applied to other domains.

Action 1c: Develop guidelines for **dataset documentation and metadata** for AI projects and AI systems. This work should include a focus on:

- Defining a minimum information standard for source description of AI data, drawing on good practices in data documentation.
- Develop guidance on how to best incorporate data provenance and lineage in metadata to improve traceability of datasets. Review work of initiatives in this field and collaborate on defining good practices and standards for this information.
- Define how IPR and licencing issues relevant to the data are presented in the documentation.

Action 1d: Develop **data characterisation documentation guidelines** and suggestions for alignment for each project or system. These guidelines would include a guidance on e.g.,:

- How to define a desired data use case for the project/system, i.e. what data is needed to reach the aims of the project/system to ensure that data selected is fit for use.
- How to identify data sensitivities, to include legal and regulatory issues relative to the use case and work to mitigate these.
- How to assess existing data for completeness (for re-users) and ensure the completeness of data that is created.
- How to undertake data improvements and manage data generated by the AI system.

Action 1e: Develop guidelines for data creators regarding the provision of **transparency for data users** around the creation and contents of the dataset, to enhance trust in these data resources and their use. This recommendation is closely related to recommendations 1c and 1d but this work will specifically focus on how to instil data users' trust in datasets they intend to use for their AI projects and systems. This work will include a focus on:

[12] See The Future Society's Review of National and International Initiatives, Summary Slide Deck for AIPR WG Meeting on Nov 18th 2020, slide 25.

- Data representativeness and coverage. Clarify whether there are issues with representativeness and coverage in the dataset, and if relevant list the steps that have been taken to eliminate bias in the dataset.
- Data accuracy and relevance. Clarify the actions that have been undertaken to verify the accuracy of the data.
- Define the legal and ethical issues that have been identified relating to the data and how have these been resolved.
- Develop trusted mechanisms (e.g., certification badges) for displaying that datasets have undergone processes that incorporate the above checks.

Recommendation 2: Underpin the creation of good quality and accessible data sources to fill data gaps in priority fields, in line with the UN Sustainable Development Goals, through targeted research and collaboration with initiatives in this field.

The focus should be on underpinning the creation of accessible and good quality data sources, according to best data governance practices. Steps should be outlined to work with governments, "AI for Social Good" initiatives, and relevant stakeholders to underpin and establish reliable data sources in priority areas. The WG should explore those areas in particular where investment is unlikely to happen, and work with other WGs and GPAI to push for action and make the data available for global benefit. As part of this work, it is important that the study also identifies gaps in dataset creation from disparate sources of data for the understanding of complex problems. One example of this is the pandemic response, where there is a lack of data sets that include socio-economic data, health record data, and genomic data leading to great risks for the public health.

Recommendation 3: Undertake research into how to improve **cross border data sharing** and write guidelines for organisations on how to address current barriers, such as:

- Intellectual Property Rights.
- Privacy and data protection legislation
- Data sovereignty

At present, there is a lack of legal certainty in relation to data access, flow and use. In other words, it is not clear who can do what with the data they are holding, or how one can acquire and analyse data in a legally compliant way. This uncertainty is even greater in relation to data sharing among jurisdictions that regulate data access and processing in a different way. The WG can identify the main areas where more certainty around AI is required and can develop concrete guidelines on how this can be achieved at a national and an international level. It can also suggest a set of best practices in preventing fragmentation among jurisdictions that can hinder data flow and availability for AI.

We stress here that the above topics should not be portrayed as only barriers, but also necessary safeguards that should be respected in data practices. Guidelines should focus on explaining how to legally and ethically undertake cross border data transfers, while research should focus on capturing good examples of how to work with legislators and policy makers to make changes that mitigate some of the challenges to international data sharing, and disseminate these more widely.

The WG should explore how to best support technological developments, such as **federated learning technologies** and **privacy-enhancing technologies** for data sharing as potential

mitigation of legal challenges, especially around personal data, and support their development and uptake where possible.

Recommendation 4: Undertake targeted research into the broad topic of **data injustice and harms** that arise from data practices around the world and identify pathways to counteract current problems. Analysis should be carried out of potential mechanisms that can overcome the challenges identified. The WG should seek out initiatives that work in this field and support them in creating concrete mechanisms to redress the harmful impacts of data in AI. We suggest priority fields to be:

- Indigenous Data Sovereignty and potential friction in relation to implementation of the FAIR principles and data openness.
- Bias in data and its impacts on society and individual rights. How to ensure inclusivity in AI data so that benefits can be more broadly realised and harms avoided.
- Environmental harms arising from data processing and storage, and how to mitigate these.
- Strengthening data capabilities in the Global South through international collaborations and networks specifically working to build soft and hard infrastructure in the region.

9. Concluding Remarks

This report has focused specifically on data and data related challenges and opportunities in the development of AI. From reviewing a variety of literature resources, we found that challenges related to the availability, accessibility and quality of data have far reaching consequences, and some may replicate or even exacerbate current inequalities. However, we also found that there are a number of initiatives that are working on solving these challenges and we have listed them here in this report, to assist the Data Governance WG and GPAI in selecting potential partners for their ongoing work.

Good data governance can mitigate and solve many of the issues highlighted in this report and it is of key importance that all stages of the data lifecycle and the AI development process are considered when developing guidelines for good data governance practices. The availability of and accessibility to good quality data for AI development is key for realising the beneficial impact AI can have across different fields and to ensure that communities and individuals are not harmed due to its application.

We are confident that the Data Governance WG and GPAI can have a real impact on strengthening data governance practices across AI fields, and see this project report as one of the steps necessary on that journey.

Resources

Abney, S., Bird, S. 2010. The Human Language Project: Building a Universal Corpus of the World's Languages. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11-16 July 2010. pp. 88–97.

Abrams, M., 2014. The Origins of Personal Data and its Implications for Governance. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2510927. Accessed 21 November 2020.

Adalbjornsson, T., 2019. Iceland's data centers are booming—here's why that's a problem. MIT Technology Review, 18 June 2019, accessed 18 October 2020.

Aidinlis, S., Smith, H., Adams-Prassl, A. and Adams-Prassl, J. (2020). Launch of a New Report: Building a Justice Data Infrastructure. [online] Oxford Law Faculty. Available at: https://www.law.ox.ac.uk/news/2020-10-08-building-justice-data-infrastructure [Accessed 26 Nov. 2020].

Andrienko, G., Andrienko, N., Giannotti, F., Monreale, A. and Pedreschi, D. (2009). Movement data anonymity through generalization. Proceedings of the 2nd SIGSPATIAL ACM GIS 2009 International Workshop on Security and Privacy in GIS and LBS - SPRINGL '09.

Alam, J.R., Sajid, A., Talib, R. and Niaz, M., 2014. A review on the role of big data in business. International Journal of Computer Science and Mobile Computing, 3(4), pp.446-453.

Allam, Z. and Dhunny, Z.A., 2019. On big data, artificial intelligence and smart cities. Cities, 89, pp.80-91.

Alston, P., 2019. Climate change and poverty: Report of the Special Rapporteur on extreme poverty and human rights. United Nations Human Rights Council, 25.

Alston, P., 2019. Report of the Special Rapporteur on extreme poverty and human rights. United Nations Human Rights Council A/74/48037.

Anonymous, 2017. Organizations Worldwide Fear Non-compliance with New European Union Data Regulation Could Put Them Out of Business. Veritas, 25 April 2017, accessed 18 October 2020.

Austin, C.A., Widyastuti, A., El Jundi, N., Nagrani, R and the RDA-COVID19 WG. (2020) COVID-19 Surveillance Data and Models: Review and Analysis, Part 1. Version 1.1, September 21, 2020, 37 pages. Available at SSRN: https://ssrn.com/abstract=3695335.

Avgerinou, M., Bertoldi, P., & Castellazzi, L., 2017. Trends in Data Centre Energy Consumption under the European Code of Conduct for Data Centre Energy Efficiency. Energies, 10(10), 1470.

Balthazar, P., Harri, P., Prater, A. and Safdar, N.M., 2018. Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics. Journal of the American College of Radiology, 15(3), pp.580-586.

Barber, D. 2012. Bayesian Reasoning and Machine Learning. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511804779.

Bartlett, J. 2018. The People vs Tech. How the internet is killing democracy (and how we can save it). Ebury Press.

Batini, C. and Scannapieco, M. (2018). DATA AND INFORMATION QUALITY : dimensions, principles and techniques.

BBC, 2018. GDPR: Data protection overhaul hits small businesses, BBC News, 22 May 2018, accessed 18 October 2020.

BBC, 2020. Google and Facebook too powerful, says watchdog. BBC News, 1 July 2020, accessed 18 October 2020.

Beauchamp, T. & Childress, J., 1985 Principles of Biomedical Ethics OUP.

Beduschi A 2020, International migration management in the age of artificial intelligence', Migration Studies.

Bennett, K.J., Olsen, J.M., Harris, S., Mekaru, S., Livinski, A.A. and Brownstein, J.S., 2013. The perfect storm of information: combining traditional and non-traditional data sources for public health situational awareness during hurricane response. PLoS currents, 5.

Bernasek, A. & Mongan, D., 2015. Our Massive New Monopolies: Amazon, Google and Facebook Have the Power to Move Entire Economies. Salon, accessed 18 October 2020.

Berne Convention for the Protection of Literary and Artistic Works, September 9, 1886, (as amended on September 28, 1979), S. Treaty Doc. No. 99-27 (1986).

Besacier, L., Barnard, E., Karpov, A., Schultz, T. 2014. Automatic speech recognition for under-resourced languages: A survey. Speech Communication. 56. Pp.85–100. DOI: 10.1016/j.specom.2013.07.008.

Boddington, P., 2017. Towards a code of ethics for artificial intelligence (pp. 27-37). Cham: Springer.

Bryson, J.J., 2018. Patiency is not a virtue: the design of intelligent systems and systems of ethics. Ethics and Information Technology, 20(1), pp.15-26.

Buolamwini J, Gebru T, 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.

Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14(0), p.2.

Carney, T., Robo-Debt Class Action Could Deliver Justice for Tens of Thousands of Australians Instead of Mere Hundreds, The Conversation (Sept. 17, 2019).

Carriere-Swallow, M.Y. & Haksar, M.V., 2019. The economics and implications of data: an integrated perspective. International Monetary Fund.

Cath, C., 2018. Governing artificial intelligence: ethical, legal and technical opportunities and challenges.

Cecco, L. 2019. 'Surveillance capitalism': critic urges Toronto to abandon smart city project. The Guardian. 6 June 2019. Available at: https://www.theguardian.com/cities/2019/jun/06/toronto-smart-city-google-project-privacy-concerns. Accessed 26 November 2020.

CERNA Committee for the Study of Research Ethics in Digital Sciences and Technologies. Allistene - Digital Sciences and Technologies Alliance. (2018). Research Ethics in Machine Learning. https://hal.archives-ouvertes.fr/hal-01724307.

Chakravorti, B. 2018. Why the Rest of the World Can't Free Ride on Europe's GDPR Rules. Harvard Business Review.

Chander, A., M. Kaminski and W. McGeveran. 2019. Catalyzing Privacy Law. Minnesota Law Review, forthcoming.

Chandran, N., 2018. 'Big tech monopolies are 'going to be a problem more and more,' media expert warns' CNBC, 11 September 2018, accessed 18 October 2020.

Chinzei, K., Shimizu, A., Mori, K., Harada, K., Takeda, H., Hashizume, M., Ishizuka, M., Kato, N., Kawamori, R., Kyo, S., Nagata, K., Yamane, T., Sakuma, I., Ohe, K. and Mitsuishi, M., 2018. Regulatory Science on AI-based Medical Devices and Systems. Advanced Biomedical Engineering, 7(0), pp.118-123.

Choroszewicz, M. and Mäihäniemi Beata (2020) "Developing a Digital Welfare State: Data Protection and the Use of Automated Decision-Making in the Public Sector Across Six Eu Countries," Global Perspectives, 1(1).

Cinnamon, J., 2017. Social injustice in surveillance capitalism. Surveillance & Society, 15(5), pp.609-625.

Clarke, R., 2019. Regulatory alternatives for AI. Computer Law & Security Review, 35(4), pp.398-409.

Coeckelbergh, M., 2019. Artificial intelligence: some ethical issues and regulatory challenges. Technology and Regulation, pp.31-34.

Cook, G., Lee, J., Tsai, T., Kong, A., Deans, J., Johnson, B. and Jardim, E., 2017. Clicking clean: Who is winning the race to build a green internet?. Greenpeace Inc., Washington, DC, 5.

Cornish, L., Jerving., S. and Ravelo, J.L. (2020) Data around COVID-19 is a mess and here's why that matters. Devex. https://www.devex.com/news/data-around-covid-19-is-a-mess-and-here-s-why-that-matters-97077.

Council of Europe. 2019. Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes'. Adopted by the Committee of Ministers on 13 February 2019 at the 1337th meeting of the Ministers' Deputies). https://search.coe.int/cm/pages/result_details.aspx?ObjectId=090000168092dd4b.

Criado Perez, C. 2019. Invisible Women: Exposing Data Bias in a World Designed for Men. New York: Abrams.

Crystal, D. 2014. Language death. Cambridge: Cambridge University Press

Cukier, K.N., Viktor Meyer-Schoenburger. 2013. The Rise of Big Data: How It's Changing the Way We Think About the World. Foreign Affairs, accessed 21 October 2020. https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data

Date, C.J. 2015. The New Relational Database Dictionary: Terms, Concepts, and Examples. Sebastopol, CA: O'Reilly Media, Inc.

Dencik, L., Hintz, A., Redden, J. and Warne, H., 2018. Data scores as governance: Investigating uses of citizen scoring in public services project report.

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [EU Database Directive 2019].

DLA Piper. 2014. IP Rights in Data Handbook, Sept. 2014.

DLA Piper. 2020. Data Protection Laws of the World, Full Handbook, accessed 16 October 2020.

DOI: https://doi.org/10.1145/2347736.2347755.

Domingos, P. 2012. A few useful things to know about machine learning. Communications of the Association for Computing Machinery. 55(10).

Environmental Protection Agency, 2007. Report to congress on server and data center energy efficiency executive summary. Public law 109-431.

Estrada-Jiménez J et al. 2019, 'On the regulation of personal data distribution in online advertising platforms', Engineering Applications of Artificial Intelligence 82, pp. 13-29.

Etzioni, A. and Etzioni, O., 2017. Incorporating ethics into artificial intelligence. The Journal of Ethics, 21(4), pp.403-418.

European Commission (2018) Turning FAIR into reality. Final Report and Action Plan from the European Commission Expert Group on FAIR data . Luxembourg Publication Office of the European Union, Luxembourg, 78 pp. https://doi.org/10.2777/1524.

European Commission (2020). 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-intelligence-feb2020_en.pdf.

European Commission for the Efficiency of Justice (CEPEJ) (2018) European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment. https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c

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

European Data Protection Board. 2020. Frequently Asked Questions on the judgment of the Court of Justice of the European Union in Case C-311/18 - Data Protection Commissioner v Facebook Ireland Ltd and Maximillian Schrems.

European Data Protection Supervisor, Declaration on Ethics and Data Protection in Artificial Intelligence (ICDPPC 2018).

European Parliamentary Research Service, EU guidelines on ethics in artificial intelligence: Context and implementation (EPRS 2019).

Felzmann, H., Villaronga, E.F., Lutz, C. and Tamò-Larrieux, A., 2019. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 6(1), p.2053951719860542.

Field, C.R., Ruskin, K.J., Benvenuti, B., Borowske, A.C., Cohen, J.B., Garey, L., Hodgman, T.P., Longenecker, R.A., King, E., Kocek, A.R., Kovach, A.I., O'Brien, K.M., Olsen, B.J., Pau, N., Roberts, S.G., Shelly, E., Shriver, W.G., Walsh, J. and Elphick, C.S. (2017). Quantifying the importance of geographic replication and representativeness when estimating demographic rates, using a coastal species as a case study. Ecography, 41(6), pp.971–981.

Findlater, L., Goodman, S., Zhao, Y., Azenkot, S. and Hanley, M. (2020). Fairness issues in AI systems that augment sensory abilities. ACM SIGACCESS Accessibility and Computing, (125), pp.1–1.

Floridi, L. and Taddeo, M. (2018) What is data ethics?, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374, 2083.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F. and Schafer, B., 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), pp.689-707

Frontier Technologies, 2020. Releasing The Power Of Digital Data For Development. [online] Available at: https://indd.adobe.com/view/883bb8ae-8702-4271-98db-86a167c01654 [Accessed 21 October 2020].

Fürber, C. (2015). Data quality management with semantic technologies. Wiesbaden: Springer Gabler.

Gaedtke, F. 2014. Can Iceland become the 'Switzerland of data'?. Al Jazeera.

Gaur, M. (2020). Privacy Preserving Machine Learning Challenges and Solution Approach for Training Data in ERP Systems. SSRN Electronic Journal.

GDPR.eu. (2020). General Data Protection Regulation. Last viewed on 22 October 2020. https://gdpr.eu/

Gianfrancesco, Milena A et al. "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data." JAMA internal medicine vol. 178,11 (2018): 1544-1547.

Ginart, A., Guan, M., Valiant, G., & Zou, J. (2019). Making AI Forget You: Data Deletion in Machine Learning. A pre-print.

Global Indigenous Data Alliance (2018) CARE Principles for Indigenous Data Governance. https://static1.squarespace.com/static/5d3799de845604000199cd24/t/5da9f4479ecab221ce848fb2/1571419335217/CARE+Principles_One+Pagers+FINAL_Oct_17_2019.pdf

Gov.UK. (2020, September 7). Open Consultation, Artificial intelligence and intellectual property: call for views. Retrieved from https://www.gov.uk/government/consultations/artificial-intelligence-and-intellectual-property-call-for-views

Greene, D., Hoffmann, A.L. and Stark, L., 2019, January. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In Proceedings of the 52nd Hawaii International Conference on System Sciences

Greenleaf, G. 2019. Global Data Privacy Laws 2019: 132 National Laws & Many Bills. 157 Privacy Laws & Business International Report, 14-18.

Grother P, Ngan M, Hanaoka, K, 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects.

Hagendorff, T., 2020. The ethics of Ai ethics: An evaluation of guidelines. Minds and Machines, pp.1-22.

Hastie, T. J., Tibshirani, R., Friedman, J. H. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: New York University Press. DOI: https://doi.org/10.1007/978-0-387-84858-7.

Hong, N. C., Cozzino, S., Genova, F., Hoffman-Sommer, M., Hooft, R., Lembinen, L., Marttila, J., Teperek, M., Ball, M., Barker, M., Berezko, O. et al. 2020. Six Recommendations for Implementation of FAIR Practice. Brussels: European Commission. DOI: 10.2777/986252
https://ec.europa.eu/info/sites/info/files/research_and_innovation/ki0120580enn.pdf.

House of Lords Select Committee on Artificial Intelligence. AI in the UK: ready, willing and able? Report of Session 2017-2019. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

Howison, J., Wiggins, A. and Crowston, K. (2011). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. Journal of the Association for Information Systems, 12(12), pp.767–797.

Huang, Y., Kuo, H.-K., Thomas, S., Kons, Z., Audhkhasi, K., Kingsbury, B., Hoory, R. and Picheny, M. 2020. Leveraging Unpaired Text Data for Training End-to-End Speech-to-Intent Systems. arXiv:2010.04284 [cs, eess]. [online] Available at: https://arxiv.org/abs/2010.04284v1 [Accessed 22 Oct. 2020].

Huyer, E. & van Knippenberg, L., 2020. The Economic Impact of Open Data. Opportunities for value creation in Europe. Brussels: European Commission.

ICO UK. 2017b. Promoting privacy with innovation within the law. News and blogs: Speech.

ICO.ORG.UK (2020). How should we assess security and data minimisation in AI? [online] Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/guidance-on-ai-and-data-protection/how-should-we-assess-security-and-data-minimisation-in-ai/ [Accessed 15 Nov. 2020].

ICRC, 2019. Mali-Niger: Climate change and conflict make an explosive mix in the Sahel. 22 January 2019, accessed 18 October 2020.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 'A Glossary for Discussion of Ethics of Autonomous and Intelligent Systems, Version 1 (2017) available at https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/eadv2_glossary.pdf.

Information Commissioner's Office (ICO) UK. 2017a. Big data, artificial intelligence, machine learning and data protection. Version 2.2.

Irvine, A., Callison-Burch, C. 2013. Combining Bilingual and Comparable Corpora for Low Resource Machine Translation. In Proceedings of the Eighth Workshop on Statistical Machine Translation, Sofia, Bulgaria, August 8-9, 2013. Pp 262–270.

International Organisation for Standardisation (1994). ISO 5725-1:1994(en) Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions.

Ishmail, N. 2018. The New Data Revolution is here. 3 August 2018, available at: https://www.information-age.com/new-data-revolution-123473921/. Accessed 26 November 2020.

Jaume-Palasí, L., Spielkamp, M. (2017). Ethics and algorithmic processes for decision making and decision support, AlgorithmWatch Working Paper No. 2.

Jones, N, 2018. 'How to stop data centres from gobbling up the world's electricity', Nature, 12 September 2018, accessed 18 October 2020.

Joseph, J. K., Dev, K. A., Pradeepkumar, A. P., & Mohan, M., 2018. Big Data Analytics and Social Media in Disaster Management. Integrating Disaster Science and Management, 287–294.

Kim, B. S., Kang, B. G., Choi, S. H., & Kim, T. G. 2017. Data modeling versus simulation modeling in the big data era: case study of a greenhouse control system. SIMULATION, 93(7), 579–594. https://doi.org/10.1177/0037549717692866.

Kim, W., Choi, B., Hong, E. et al. A Taxonomy of Dirty Data. Data Mining and Knowledge Discovery 7, 81–99 (2003). https://doi.org/10.1023/A:1021564703268.

Klein, A., 2019. 'Credit denial in the age of AI' Brookings Institute, accessed 18 October 2020.

Kroll, J.A. (2018). The fallacy of inscrutability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376: 20180084. http://dx.doi.org/10.1098/rsta.2018.0084.

Kukutai, T., Taylor, J. 2016. Indigenous Data Sovereignty: Toward an agenda. ANU Press, Australia. DOI: 10.22459/CAEPR38.11.2016 .

LeCun, Y. (2019). The Next AI Revolution "Will Not Be Supervised" Because Not Everything Can Be Predicted. [online] Eyerys. Available at: https://www.eyerys.com/articles/people/1560388243/opinions/the-next-ai-revolution-will-not-be-supervised [Accessed 23 Oct. 2020].

Lepri, B., Oliver, N., Letouzé, E., Pentland, A. and Vinck, P. (2017). Fair, Transparent, and Accountable Algorithmic Decision-making Processes. Philosophy & Technology, 31(4), pp.611–627.

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

Levy, S., 2020. Facebook: how Mark Zuckerberg gobbled up Instagram and Whatsapp in the battle for your data. The Sunday Times, 23 February 2020, accessed 18 October 2020.

Liu, Y.-N., Li, J.-Z. and Zou, Z.-N. (2016). Determining the Real Data Completeness of a Relational Dataset. Journal of Computer Science and Technology, 31(4), pp.720–740.

Lomas, N., 'UK health minister sets out tech-first vision for future care provisio', TechCrunch (2018) https://techcrunch.com/2018/10/17/uk-health-minister-sets-out-tech-first-vision-for-future-care-provision/

Loshin, D. (2009). Data Quality and MDM. Master Data Management, pp.87–103.

Lucivero, F., 2020. Big data, big waste? A reflection on the environmental sustainability of big data initiatives. Science and engineering ethics, 26(2), pp.1009-1030.

Luengo-Oroz, M, Pham, K.H., Bullock, J., Kirkpatrick, R., Luccioni, A., Rubel, S., Wachholz, C., Chakchouk, M., Biggs, P., Nguyn, T., Purnat and Mariano, B. (2020) Artificial intelligence cooperation to support the global response to COVID-19. Nature Machine Intelligence 2, 295-297(2020. DOI: https://doi.org/10.1038/s42256-020-0184-3

Lum, K. and Isaac, W., 2016. To predict and serve? Significance, 13(5), pp.14-19.

Martens, B. 2018 The Importance of Data Access Regimes for Artificial Intelligence and Machine Learning. JRC Digital Economy Working Paper. http://dx.doi.org/10.2139/ssrn.3357652

Mercer, S. T. 2020. The Limitations of European Data Protection as a Model for Global Privacy Regulation. 114 AJIL UNBOUND 20.

Metz, C., 2019. AI is learning from humans. Many humans. The New York Times, 19 August 2019, accessed 21 October 2020.

Mitchell, M. and Caudy, S.M. (2013) Examining Racial Disparities in Drug Arrests.Justice Quarterly 2: 288–313.

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016). The ethics of algorithms: Mapping the debate, Big Data & Society, July – December, 1–21.

Moore, M., 2016. The Giants and Civic Power, CMPCP, Policy Institute at King's College London, pp. 5-20. https://doi.org/10.18742/pub01-027.

National Science and Technology Council. 2019. National Artficial Intelligence Research and Development Strategic Plan: 2019 Update. Select Committee on Articial Intelligence Report. June. https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. 2016. Computational sociolinguistics: a survey. Computational Linguistics. 42(3), pp. 537–593.

Noble, S. U. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. New York: New York University Press; Sap, M., Card, D., Gabriel, S., Choi, Y., Smith N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. August, 2019. Florence: Association for Computational Linguistics. pp. 1668-1678. DOI: 10.18653/v1/P19-1163.

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

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T. and Staab, S., 2020. Bias in data-driven artificial intelligence systems−An introductory survey. WIREs Data Mining and Knowledge Discovery, 10(3).

Nunes, M. L., Pereira, A. C., & Alves, A.C., 2017. Smart products development approaches for Industry 4.0. Procedia Manufacturing, 13, 1215–1222.

O'Leary, D. E. 2013. Artificial Intelligence and Big Data. IEEE Intelligent Systems, 28(2), pp. 96-99. DOI: 10.1109/MIS.2013.39.

Obermeyer, Z. etal. (2019), "Dissecting racial bias in an algorithm used to manage the health of populations", Science, Vol.366/6464, pp.447-453, abstract.

Open Geospatial Consortium (2012), OGC® Coverage Implementation Schema.

OECD, (2019a) Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies. https://www.oecd-ilibrary.org/sites/276aaca8-en/1/2/2/index.html?itemId=/content/publication/276aaca8-en&_csp_=a1e9fa54d39998ecc1d83f19b8b0fc34&itemIGO=oecd&itemContentType=book#section-d1e1463

OECD, OECD Legal Instruments (2019b) Recommendation of the Council on Artificial Intelligence OECD/LEGAL/0449 Adopted on 22/05/2019, 2.1.(b) Investing in AI research and development https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OECD, Background paper for the G20 AI Dialogue, Digital Economy Task Force (2020a) https://www.oecd.org/health/trustworthy-artificial-intelligence-in-health.pdf p 13

OECD, OECD Policy Responses to Coronavirus (COVID-19), (2020b) Why open science is critical to combatting COVID-19. https://www.oecd.org/coronavirus/policy-responses/why-open-science-is-critical-to-combatting-covid-19-cd6ab2f9/

OECD, OECD Policy Responses to Coronavirus (COVID-19), (2020c) Using artificial intelligence to help combat COVID-19. https://www.oecd.org/coronavirus/policy-responses/using-artificial-intelligence-to-help-combat-covid-19-ae4c5c21/

OECD.AI (2019). Investing in AI research and development (OECD AI Principle) - OECD.AI. [online] Oecd.ai. Available at: https://oecd.ai/dashboards/ai-principles/P10 [Accessed 26 Nov. 2020].

OECD.AI (2020), visualisations powered by JSI using data from MAG, accessed on 3/3/2020, www.oecd.ai.

Oswald, M., Grace, J., Urwin, S., Barnes G.C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality, Information & Communications Technology Law, 27:2, 223-250, DOI: 10.1080/13600834.2018.1458455.

Panda, S.S. and Jena, D. (2020). Decentralizing AI Using Blockchain Technology for Secure Decision Making. Algorithms for Intelligent Systems, pp.687–694.

Panel for the Future of Science and Technology of the European Parliamentary Research Service (PFST). (2019). A governance framework for algorithmic accountability and transparency. Scientific Foresight Unit (STOA) PE 624.262 – April 2019. https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf.

Panel for the Future of Science and Technology of the European Parliamentary Research Service (PFST). (2020). The ethics of artificial intelligence: Issues and initiatives. Scientific Foresight Unit (STOA) PE 634.452 – March 2020. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf.

Pang, W. (2019). How to Ensure Data Quality for AI. [online] insideBIGDATA. Available at: https://insidebigdata.com/2019/11/17/how-to-ensure-data-quality-for-ai/ [Accessed 23 Oct. 2020].

Perera, H., Hussain, W., Mougouei, D., Shams, R. A., Nurwidyantoro, A., & Whittle, J., 2019. Towards Integrating Human Values into Software: Mapping Principles and Rights of GDPR to Values. 2019 IEEE 27th International Requirements Engineering Conference (RE).

Perlin, M. (2020). Quality Assurance for Artificial Intelligence. [online] Medium. Available at: https://towardsdatascience.com/quality-assurance-for-artificial-intelligence-d935fc6b238.

Phillips P., Grother P, Michaels R, Blackburn D., Tabassi E., Bone M., 2003. Face Recognition Vendor Test 2002

Pinto Dos Santos, D., & Baeßler, B., 2018. Big data, artificial intelligence, and structured reporting. European radiology experimental, 2(1), 42. https://doi.org/10.1186/s41747-018-0071-4

Ray, T. (2020) AI runs smack up against a big data problem in COVID-19 diagnosis. April 5, 2020. ZDNET. https://www.zdnet.com/article/ai-runs-smack-up-against-a-big-data-problem-in-covid-19-diagnosis/

Reed, C., 2018. How should we regulate artificial intelligence?. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2128), p.20170360

Regan, P 1995, Legislating Privacy, Chapel Hill, NC: University of North Carolina Press.

Reuters (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Last viewed 22 October 2020. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Richardson, Rashida and Schultz, Jason and Crawford, Kate, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice (February 13, 2019). 94 N.Y.U. L. REV. ONLINE 192 (2019), Available at SSRN: https://ssrn.com/abstract=3333423

Roberts, R., 2009. Virtual Research Environments. Oxford, Chandos Publishing. P178-179. doi.org/10.1016/B978-1-84334-562-6.50025-6

Rodríguez Rojas, L.A., Cueva Lovelle, J.M., Tarazona Bermúdez, G.M. and Montenegro, C.E., 2013. Open Data as a key factor for developing expert systems: a perspective from Spain.

Rosenblum, P., Maples, S. and Revenue Watch Institute (2009). Contracts confidential : ending secret deals in the extractive industries. New York, Ny: Revenue Watch Institute.

Ross, M.E, A.R. Kreider, Y.S. Huang, M. Matone, D.M. Rubin, A.R. Localio, 2015. Propensity Score Methods for Analyzing Observational Data Like Randomized Experiments: Challenges and Solutions for Rare Outcomes and Exposures, American Journal of Epidemiology, 181.12. P989–995, https://doi.org/10.1093/aje/kwu469

Rovatsos, M., Luger, E. 2020. Data Ethics, AI and Responsible Innovation. Last viewed on 22 October 2020. https://www.wiki.ed.ac.uk/display/DEARI

Rowe, M & R Muir 2019, 'Big Data Policing: Governing the Machines?' Policing and Artificial Intelligence, Taylor & Francis, London: UK.

Samarati, P., Sweeney, L.: Generalizing Data to Provide Anonymity when Disclosing Information (Abstract). In: Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, p. 188 (1998)e.com/document/d/19DyV75O-uD-jBtZ7TGciUGOO50j94c83nJXT5kGMmdQ/edit#

Santosh, K.C. (2020) AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data. Nature Public Health Emergency Collection. J Med Syst. 2020; 44(5):93 doi: 10.1007/s10916-020-01562-1

Sap, M., Card, D., Gabriel, S., Choi, Y., Smith N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. August, 2019. Florence: Association for Computational Linguistics. pp. 1668-1678. DOI: 10.18653/v1/P19-1163.

Sap, M., Card, D., Gabriel, S., Choi, Y., Smith N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. August, 2019. Florence: Association for Computational Linguistics. pp. 1668-1678. DOI: 10.18653/v1/P19-1163.

Sardi, S., Vardi, R., Meir, Y. et al., 2020. Brain experiments imply adaptation mechanisms which outperform common AI learning algorithms. Sci Rep 10, 6923. https://doi.org/10.1038/s41598-020-63755-5

Sartor, G. and F. Lagioia. The impact of the General Data Protection Regulation (GDPR on artificial intelligence. (2020). European Parliamentary Research Service, Scientific Foresight Unit, PE 641.530.

Scelta, G., Rashid, H., Cheng, H. W. J., LaFleur, M., Parra-Lancourt, M., Julca, A., Hunt, N., Islam, S., Kawamura, H. 2019. Frontier Technology Quarterly January 2019: Data Economy - radical transformation or dystopia?, pp.3. https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/publication/FTQ_1_Jan_2019.pdf

Schwalbe, N. and B. Wahl. 2020. Artificial Intelligence and the future of global health. The Lancet, Review. (2020) Volume 395, Issue 10236.

Sciforce. (2019, October 11). NLP for Low-Resource Settings (Blog post). Retrieved from https://medium.com/sciforce/nlp-for-low-resource-settings-52e199779a79

Sebastian-Coleman, L. (2013). DQAF Concepts. Measuring Data Quality for Ongoing Improvement, pp.57–69.

Silberg, J and Manyika, J, 2019. Notes from the AI frontier: Tackling bias in AI (and in humans)

Simperl, E. K. O'Hara and R. Gomer. Analytical Report 3: Open Data and Privacy. 2020. European Data Portal.

Son H.S. (2020) Validity Evaluation for the Data Used for Artificial Intelligence System. In: Bi Y., Bhatia R., Kapoor S. (eds) Intelligent Systems and Applications. IntelliSys 2019. Advances in Intelligent Systems and Computing, vol 1037. Springer, Cham. https://doi.org/10.1007/978-3-030-29516-5_28

Stevenson, M. and Mayson, S.G. (2018) The Scale of Misdemeanor Justice. Boston University Law Review 98 (731): 769–770.

Stoltzfus, J. (n.d.). How do chatbots deal with accents? [online] Techopedia.com. Available at: https://www.techopedia.com/how-do-chatbots-deal-with-accents/7/33229 [Accessed 26 Nov. 2020].

Stork, D.G., 2000, June. Open data collection for training intelligent software in the open mind initiative. In Proceedings of the Engineering Intelligent Systems (EIS2000)

Strickland,E. (2019), How IBM Watson Overpromised and Underdelivered on AI Health Care -IEEE Spectrum, IEEE Spectrum, https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K., Wang, W. Y. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp.1630–1640. DOI: 10.18653/v1/P19-1159.

Strong, D.M., Lee, Y.W. and Wang, R.Y. (1997). Data quality in context. Communications of the ACM, 40(5), pp.103–110.

Taylor, A., 2018. Failover architectures: The infrastructural excess of the data centre industry—Failed architecture. Failed Architecture.

Taylor, P. 2020. Short Cuts: Ofqual and the Algorithm. London Review of Books. 42(17). London: LRB.

Toda, T., Kawai, H., and Tsuzaki, M. (2004). Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing 2004.

Toews, R., 2020. The Next Generation Of Artificial Intelligence. Forbes. Retrieved from https://www.forbes.com/sites/robtoews/2020/10/12/the-next-generation-of-artificial-intelligence/

TRIPS: Agreement on Trade-Related Aspects of Intellectual Property Rights. 1994. Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 3; 33 I.L.M. 1197. Apr. 15, 1994.

Trueman, C. 2019. Why data centres are the new frontier in the fight against climate change. Computer World. 9 August 2019. Available at: https://www.computerworld.com/article/3431148/why-data-centres-are-the-new-frontier-in-the-fight-against-climate-change.html. Accessed 26 November 2020.

U.S Supreme Court Decision, Feist Publications, Inc., v. Rural Telephone Service Co., 499 U.S. 340 (1991).

UN General Assembly. 2013. The right to privacy in the digital age, resolution adopted by the General Assembly, 18 December 2013, A/RES/68/167.

Ulhaq, A., and Burnmeister, O. COVID-19 Imaging Data Privacy by Federated Learning Design: A Theoretical Framework. Preprint at https://arxiv.org/pdf/2010.06177.pdf (2020).

United Nations (IEAG) Independent Expert Advisory Group on a Data Revolution for Sustainable Development., 2014. A World that Counts Mobilising the Data Revolution for Sustainable Development.

United Nations, 2015. Transforming our world: The 2030 agenda for sustainable development (working papers). eSocialSciences.

United Nations Development Group, 2017. Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda (UNDG 2017)

Vaughan, A., 2015. How viral cat videos are warming the planet. The Guardian, 25 September 2015.

Veale, M., Binns, R., Edwards, L. 2018. Algorithms that remember: model inversion attacks and data protection law. In Philosophical Transactions of the Royal Society A. 376:20180083. DOI: http://doi.org/10.1098/rsta.2018.0083

Vigdor, N. (2019, November 10). Apple Card Investigated After Gender Discrimination Complaints. Retrieved from https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html .

Vinuesa, R. et al. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications, (2020) 11:233 | https://doi.org/10.1038/s41467-019-14108-y | www.nature.com/naturecommunications

VoPham, T., Hart, J.E., Laden, F. and Chiang, Y.-Y. (2018). Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. Environmental Health, 17(1).

Wagner, B. (2018). Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping? In Hildebrandt, M., Being Profiling. Cogitas ergo sum. Amsterdam University Press.

Wakchaure A., Eaglin R. and Motlagh B., 2008. A technique for the quantitative measure of data cleanliness, 2008 IEEE Conference on Cybernetics and Intelligent Systems, Chengdu, 2008, pp. 1258-1263, doi: 10.1109/ICCIS.2008.4670930.

Wakefield, J. (2020) Coronavirus: AI steps up in battle against Covid-19. 17th April 2020. BBC News. https://www.bbc.co.uk/news/technology-52120747

Wang, R. and Strong, D., 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), pp.5-33.

Wang, S. Y., Pershing, S., Lee, A. Y., & AAO Taskforce on AI and AAO Medical Information Technology Committee, 2020. Big data requirements for artificial intelligence. Current opinion in ophthalmology, 31(5), 318–323. https://doi.org/10.1097/ICU.0000000000000676

Weinberg, J., Freese, J. and McElhattan, D., 2014. Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample. Sociological Science, 1, pp.292-310.

Whitehead, B., Andrews, D., Shah, A., & Maidment, G., 2014. Assessing the environmental impact of data centres part 1: Background, energy use and metrics. Building and Environment,82, 151–159.

World Economic Forum, 2014. Rethinking Personal Data: A New Lens for Strengthening Trust. Prepared in collaboration with A.T. Kearney. Available at https://www.weforum.org/reports/rethinking-personal-data. Accessed 18 November 2020.

World Health Organisation, (2015) WHO consultation on Data and Result Sharing During Public Health Emergencies: Background Briefing. Authors: Ben Goldacre, Sian Harrison, Kamal R. Mahtani and Carl Henegan. Centre for Evidence-Based Medicine, University of Oxford, https://www.who.int/medicines/ebola-treatment/background_briefing_on_data_results_sharing_during_phes.pdf?ua=1

Wiggers, K. 2019. Mozilla updates Common Voice dataset with 1,400 hours of speech across 18 languages. Retrieved from https://venturebeat.com/2019/02/28/mozilla-updates-common-voice-dataset-with-1400-hours-of-speech-across-19-languages/ .

Wikipedia, Facebook–Cambridge Analytica data scandal. Viewed on 22 October 2020. Retrieved from https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Wilks, J. 2014. IP rights in Data Handbook: Protecting and exploiting IP in data, big data and databases internationally. DLA Piper. September 2014.

Williams, E., 2011. Environmental effects of information and communications technologies. Nature,479(7373), 354–358.

Yoon, J., Drumright, L.N. and van der Schaar, M. (2020). Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). IEEE Journal of Biomedical and Health Informatics, 24(8), pp.2378–2388.

Zhou, Y., Wang, F., Tang, J., Nussinov, R. and Cheng, F. 2020 Artificial intelligence in COVID-19 drug repurposing. Lancet Digital Health, https://doi.org/10.1016/S2589-7500(20)30192-8

Zuboff, S., 2015. Big other: surveillance capitalism and the prospects of an information civilization. Journal of Information Technology, 30(1), 75–89.

Zuboff, S., 2019. Shoshana Zuboff: Facebook, Google and a dark age of surveillance capitalism. Financial Times, 25 January 2019, accessed 18 October 2020.

Zuiderwijk, A., Gaseó, M., Parycek, P. and Janssen, M., 2014. Special issue on transparency and open data policies: Guest editors' introduction. Journal of theoretical and applied electronic commerce research, 9(3), pp.I-IX.

Zwijnenburg, W., et al, 2020. Solving the jigsaw of conflict-related environmental damage: Utilizing open-source analysis to improve research into environmental health risks. Journal of Public Health, 42(3), 352–360

Annex A- Initiatives and projects working on challenges relating to data governance, availability and accessibility.

Please note that this list is not exhaustive but consists of resources and tools that were discovered in the course of the literature review and in consultations with the experts of the Data Governance Working Group.

Data gaps and data availability

CLARIN - European Research Infrastructure for Language Resources and Technology has the mission to create and maintain an infrastructure to support the sharing, use and sustainability of language data. CLARIN currently offers 12 types of corpora, lexical resources and language data tools. https://www.clarin.eu/

The Lacuna Fund is working to fill data gaps in the following fields: Language, Agriculture and Health, to assist data scientists, researchers and social entrepreneurs in Lower- and Middle-Income Countries with access to data. https://lacunafund.org/

Mozilla Common Voice is part of Mozilla's initiative to make more voice data available for AI development. Anyone can submit a recording of their voice and through this crowdsourcing initiative thousands of hours of speech for 60 different languages have been collected. https://commonvoice.mozilla.org/en/about

Data catalogues and repositories

FAIRsFAIR project is currently working on practical solutions to boost the impact of data repositories and make them enable FAIR data to boost the findability of both repositories and their data. https://www.fairsfair.eu

The Global Data Access Framework ' aims at leveraging the revolution in advanced analytics and Artificial Intelligence to support the achievement of the UN Sustainable Development Goals (SDGs). It has been envisioned as a precursor for the AI for SDGs Center (AI4SDG) and is a part of the AI Commons initiative https://thefuturesociety.org/2019/11/15/global-data-access-framework-gdaf/

RDA Research Data Repository Interoperability WG is working on establishing standards for interoperability between different repository platforms.
https://www.rd-alliance.org/groups/research-data-repository-interoperability-wg.html

UN International Telecommunications Union (ITU) has set up a global AI' repository to identify AI related projects, research initiatives, think-tanks and organizations that can accelerate progress towards the UN SDGs. https://www.itu.int/en/ITU-T/AI/Pages/ai-repository.aspx

World Wide Web Consortium (W3C) Data Catalogue Vocabulary (DCAT) DCAT enables a publisher to describe datasets in a catalog using a standard model and vocabulary This can increase the discoverability of datasets and makes federated search for datasets across catalogs in multiple sites possible using the same query mechanism and structure. https://www.w3.org/TR/vocab-dcat-2/

Data documentation and provenance

Data Documentation Initiative (DDI) has developed the DDI - Cross Domain Integration

specification intended to help with data integration across domain and institutional boundaries. DDI-CDI will be able to describe data and its provenance at a detailed, machine-actionable level. The DDI-CDI specification is currently under public review. https://ddialliance.org/announcement/public-review-ddi-cross-domain-integration-ddi-cdi

RDA Research Data Provenance IG is working on recommendations regarding frameworks for documenting data transactions and how to account for modifications and how to assess data quality. https://www.rd-alliance.org/groups/research-data-provenance.html

World Wide Web Consortium (W3C) Provenance WG developed the PROV data model for provenance interchange on the Web. All documents, along with the model can be found at https://www.w3.org/TR/prov-overview/

Data governance

CODATA International Data Policy Committee provides expert input on the development and implementation of data policies to a range of international initiatives. A part of the committee's strategy is to support implementation of data principles and practices. https://codata.org/initiatives/strategic-programme/international-data-policy-committee/

Expert Advisory Group on Data Access report on Governance of Data Access (EAGDA) was a group set up by the Wellcome Trust to good working practices for managing and using data from cohort studies. It ran between 2012-2017 and produced various guidelines about terms for data use, sanctions and accountability, data management plans, infrastructure development and data curation, governance of data access, risks of harm from data misuse, incentives to support data access, protecting confidentiality for research participants. https://cms.wellcome.org/sites/default/files/governance-of-data-access-eagda-jun15.pdf

RDA International Indigenous Data Sovereignty Interest Group is a Research Data Alliance group that aims to promote indigenous data sovereignty (ID-Sov), which is defined as 'the right of a nation to govern the collection, ownership, and application of its own data'. There are three operating subgroups: Mana Raraunga - Maori Data Sovereignty Network, the United States Indigenous Data Sovereignty Network (USIDSN), and the Maiamnayri Wingara Aboriginal and Torres Strait Islander Data Sovereignty Group in Australia. https://www.rd-alliance.org/groups/international-indigenous-data-sovereignty-ig

Legal interoperability

CODATA-RDA Interest Group on Legal Interoperability of Research Data has investigated issues related specifically to IPR of data and developed a set of principles and practical implementation guidelines. https://codata.org/initiatives/working-groups/legal-interoperability/

Data Quality

IEEE Standards Association is currently working on developing The Ethics Certification Programme for Autonomous and Intelligent Systems (ECPAIS) to create 'specifications for certification and marking processes that advance transparency, accountability and reduction in algorithmic bias in Autonomous and Intelligent Systems.' This work, although focused on the systems, could serve as a starting point for a similar process for AI data. https://standards.ieee.org/industry-connections/ecpais.html

WDS[13]*/RDA* Assessment of Data Fitness for Use WG worked on an assessment criteria for data fitness for use, including examining the potential for a certification in this respect. https://www.rd-alliance.org/groups/assessment-data-fitness-use

Data management tools and resources

DMPonline is an open source, online tool DMPRoadmap codebase, which is jointly developed by the Digital Curation Centre (DCC) and the University of California Curation Center (UC3). The DCC & UC3 work closely with research funders and universities to produce a tool that generates active DMPs and caters for the whole lifecycle of a project, from bid-preparation stage through to completion. https://dmponline.dcc.ac.uk/

Open Science Framework, is an open source, online tool that aims to increase transparency and reproducibility in research. Researchers can use the tool to plan for data management and record all data, code, software, tests, etc used for their studies. This information is shared among researchers who use the tool. https://help.osf.io/

FAIR data

EOSC FAIR Working Group provides the European Open Science Cloud (EOSC) with recommendations for the implementation of Open and FAIR practices. https://www.eoscsecretariat.eu/working-groups/fair-working-group

FAIRsFAIR project aims to supply practical solutions for the use of the FAIR data principles throughout the research data life cycle. It is tasked with providing a platform for EOSC data providers and repositories, as well as rules for participation in EOSC projects. https://fairsfair.eu/

FAIR4Health project aims to facilitate and encourage the EU health research community to FAIRify, share and reuse. One of the objectives is 'to develop and validate intuitive, user-centered technological tools to enable the translation from raw (meta)data to FAIR (meta)data and support the FAIRification workflow, i.e., the FAIR4Health Platform and Agents'. https://www.fair4health.eu/

FAIRSharing is a resource for researchers. It provides metadata standards, collections of data, databases as well as education on data policies. The distinctive characteristic of their data resources is that they comply with the FAIR principles. https://fairsharing.org/

[13] ICSU World Data System