Community Coordinated Artificial Intelligence

Towards a unified framework for the democratisation of AI

Anders Jakob Sivesind

Thesis submitted for the degree of
Master in Informatics: Programming and System Architecture
60 credits

Institute for Informatics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2021

Community Coordinated Artificial Intelligence

Towards a unified framework for the democratisation of AI

Anders Jakob Sivesind

Community Coordinated Artificial Intelligence

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

Contributing to an emerging AI-paradigm shift, this thesis presents a unified socio-technical framework called Community Coordinated Artificial Intelligence (CoCoAI), which expands the horizons of the AI expertocracy. Currently, AI is used mostly by companies (or governments) to analyse people's behaviour to serve their own commercial interests. I argue instead how people, not companies, could ultimately benefit from the development and use of AI. To achieve this goal, I have established four research objectives. The first research objective is performing a literature review on the democratisation of AI. This serves as the scientific foundation for my thesis. My second objective is to establish a definition that unifies the various understandings of what the topic entails. Further, my third objective is to create an overview of the challenges and solutions to the democratisation of AI presented in the literature. Finally, my fourth research objective is to develop a socio-technical framework for the democratisation of AI, using the definition, challenges and solutions I established in my previous objectives.

To form the scientific foundation necessary to accomplish this work, I will perform a structured configurative review of the literature on the topic. By creating a unified definition and an overview of the challenges and solutions, I will establish a foundation for further research on the topic.

Moreover, my framework can inform the design of AI platforms and projects, promoting processes that ensure democratic control of the technology. CoCoAI provides benefits on three levels. On a societal level, CoCoAI promotes AI solutions that are beneficial to society as a whole, protecting rights and democratic values, and avoiding solutions that discriminate against social groups, or otherwise treats them unfairly. For organisations, CoCoAI increases the availability of AI resources, technology and expertise. This can enable more organisations to benefit from AI for their use case. Finally, on the individual level, CoCoAI promotes education, knowledge sharing, transparency and beneficial solutions. Increased access to educational resources and knowledge sharing in AI can contribute to a society where more people have a basic understanding of the technology. By also having more transparency surrounding the AI systems in use, users will be able to make more informed decisions in their interactions with such AI services and systems. Finally, CoCoAI promotes access to beneficial AI solutions, both by making corporate AI development processes more democratic, but also by enabling the creation of more grassroots AI projects as a result of better access to AI resources, knowledge and technology.

Acknowledgements

Writing this thesis has been a fascinating dive into understanding AI as more than a mere technology, but in fact as a trans-disciplinary field of research. This is a core challenge for research on information technologies in general, which involves not only technical challenges such as programming, but also understanding the socio-technical dynamics and ethical implications of the problem. This involves interactions with users, organisations, governments and trans-national institutions. Throughout the process of writing this thesis, I have received invaluable support and guidance from co-students, lecturers and supervisors.

First and foremost, I would like to thank my supervisor Christian Johansen for opening many doors for me throughout the process of this thesis project. Attending various AI related conferences was an invaluable source inspiration and insights into the more academic aspects of computer science. In particular, the AI summer school and conference in Oxford will remain with me as an important memory. Further, the opportunity for me to attend several meetings, participate in a workshop and present my work on several occasions in relation to the IoTSec and SCOTT research projects, provided me with important experience and feedback that I believe will be very useful for the future. I also greatly appreciate the advice offered and the interesting discussions we have had throughout the project.

I am also very grateful to Adam Zachary Wyner for hosting me at Swansea University, while we worked together for several months in the beginning of my thesis project. I wish to thank Tore Pedersen for many interesting discussions, useful inputs and literature tips. Thank you to Johanna Johansen for notifying me about relevant literature and papers. I also want to thank Clara Julia Reich for many interesting discussions and valuable input. Finally, I wish to thank my family for all their wonderful support.

Oslo, May, 2021

Anders Jakob Sivesind

Contents

List of Tables

List of Figures

1 Introduction

AI is a technology in great demand, but there is a very small supply of experts with the knowledge and experience necessary to develop AI capabilities usable for solving real-world problems. In addition to this, building datasets big enough to solve real problems is a complicated, tedious and expensive process. Consequently, the development and implementation is mostly being done by companies and governments with the funds to pay for the very high costs.

There are a growing number of people and organisations who have a strong wish to start using AI to help solve their problems, but they lack the expertise or funds necessary to take on such a project (Allen et al., 2019; Kobayashi et al., 2019). Furthermore, AI seems to have a large number of potential benefits, such as image recognition, voice recognition or sentiment analysis. However, these potential benefits are in stark contrast with the potential for misuse of the same technology to gain influence over people and, indeed, to manipulate them in numerous ways (e.g. O'Neil, 2016). Moreover, AI has brought a massive difference in power between those that develop and use AI and the lay people, even more so when they are in the scope of the AI (Hall, 2017; Jiang et al., 2017; R. Malhotra and D. K. Malhotra, 2003; Manheim and Kaplan, 2019). I am particularly concerned with how AI can remove any sense of privacy, even without the person's awareness of relinquishing such a basic right. Thus, in this thesis I will often focus on examples related to privacy.

1.1 AI from a technical perspective

In this thesis I will interpret AI through a definition by Schalkoff (1990): 'A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes'. Thus, by extrapolation, AI is a computational process that emulates intelligent behaviour. This is intentionally a very broad interpretation of the concept, which enables the inclusion of traditional rule-based AI, classical machine learning algorithms, neural networks and other 'intelligent' algorithms. However, variations of machine learning may be the most relevant type of AI in relation to a number of topics throughout the thesis.

When looking at AI from a technological perspective, the primary function is to extract information from complex data automatically. This fundamental property implies two critical abilities:

1. the ability to automatically extract information that a person would not be able to see otherwise, because the information is, so to speak, 'hidden' in the usually very large amount of data;

2. the ability to obtain information at a much greater scale than a person would be able to, because of the speed with which machine learning algorithms can process large amounts of complex data.

AI can be divided into two major categories: rule-based AI, and machine learning (ML). Rule-based AI is a type of AI that is designed and

programmed by an expert human. It uses rules tailored to detect specific features in data and to make decisions based on these features. Machine learning, on the other hand, learns behaviour directly from observing a large amount of data, and looking for statistical features in the dataset that enable it to make decisions from new data.

ML involves three technical elements: (A) an algorithm able to learn from examples, (B) a dataset containing the data from which the algorithm will learn, and (C) a computing system able to execute the algorithm on the dataset. There are numerous algorithms from which to choose, each of which offers technical properties that differ somewhat from those of the others, but all share the core purpose of extracting information from mostly unstructured data. Further, the dataset needs to contain a significant amount of data that may require substantial manual labour to gather and compile. Finally, the computing system must be able to run the learning algorithm. For simpler models, the hardware does not have to be exceptionally powerful, but the more powerful it is, the more sophisticated the algorithms it will be able to run and the faster it will be able to complete the learning, which in turn enables the processing of larger datasets in a reasonable amount of time.

1.2 Corporate AI

Facebook, Google, Spotify, Microsoft, Amazon and Snapchat are all companies associated with the modern digital society, offering services that enable people to communicate and stay connected all the time, everywhere. In the use of these services, people leave behind copious amounts of data about themselves. Sometimes the data is given deliberately, such as entering the date of birth when registering on a website. Often, we provide the data voluntarily, but perhaps not consciously, such as when buying a book from Amazon. At other times, the companies extract information from the patterns in people's usage of a service, such as likes on Facebook (Kosinski et al., 2013).

People are familiar with the dynamics of the traditional economy. They spend money on some service or product, and then they can earn it back through work. But the new types of services represent another set of dynamics: surveillance capitalism (Zuboff, 2019). Most people may not entirely understand or appreciate how these dynamics work. A key difference from traditional economic systems is that, once people have shared information about themselves, there is no way to unshare it – unlike with money.

According to EU regulations (European Parliament, 2016), people have the right to delete the information a company has about them. But if control of data is lost, such as, if the company gets hacked, or if the data is sold, then it is too late. Perhaps even more concerning is if people do not understand under which terms they are providing their data, then how can they know that they agree with how it is being used and shared in the first place?

The quote 'If you are not paying for it, you're not the customer; you're the product being sold' (Lewis, 2010) draws attention to another critical misconception. Google's and Facebook's services are not free; people don't pay for them with their money, but with information about themselves. This

realisation raises another question: why is their data so valuable?

Surveillance capitalism is the economic system underlying many of the services people use on a daily basis (Zuboff, 2019). Instead of being a deal between two actors, e.g., trading money for a service, it is a deal between many actors, wherein the user provides data about themselves, rather than money, and that data is later used to generate money for the service provider. The services seem free because people do not have to pay money for them, and people seem not to have the same feeling of value for their data as they have for their money. It appears quite natural that people do not have an inherent sense of this value, as it is very abstract; i.e., technical, sociological and economic knowledge is required to appreciate the importance of data.

Surveillance capitalism thrives also because of another aspect which is even more difficult for lay people to comprehend or manage, and that is, since the value of information is linked to technology and context, a piece of information that was considered trivial at one point may become critical sometime later, or otherwise be a piece in a puzzle with other data that may enable someone to extract more personal information. The importance of information is determined not only by its present but also by its future value, sometimes extracted by more sophisticated algorithms (Schneier, 2015).

Personalisation is arguably the most natural use of personal information. A company may collect someone's name so that they can address them by their first name, to make their service feel more personal. They may note who the closest friends are, so that their posts appear at the top of the social feed. Maybe the company uses a person's interest in American politics to determine that they may wish to see the latest news from the American election.

Personal data also lends itself to information extraction, as does any other type of data with which an AI system works: e.g., looking at the statistical patterns in large quantities of data and tracking who buys what, who talks to whom or which news one reads. With enough time and data, it is possible to build accurate algorithms that can predict a person's behaviour, beliefs or values (Kosinski et al., 2013). A problematic consequence of such algorithms is that these companies may know more about people's private lives than do their closest friends, family, or even themselves (Peters, 2019).

Combining these two uses, a company can deploy the previously extrapolated information to personalise a strategy that targets a person individually. The strategy takes into account what the person cares about and precisely what matters in their decision process, orchestrating everything to tip their decisional scale (Matz et al., 2017). Cambridge Analytica was famously involved in the 2016 American presidential election, where they explored the use of this technology (Isaak and Hanna, 2018). There has been some discussion as to whether they made a significant impact on the outcome or not, but studies have shown that psychological targeting can have a real impact on people's actions (Matz et al., 2017). I will, however, argue that it does not matter to which degree the political micro-targeting affected the results, as the mere attempt violates citizens' fundamental rights to privacy and freedom (André et al., 2018; Floridi et al., 2018; Thwaite, 2019). While the Cambridge Analytica scandal received a lot of attention in recent

years, it is far from the first attempt at using AI to affect election results. A quick look back in the history of U.S. elections reveals examples such as the presidential election campaign of Obama, where Google participated in the campaign, and that of John F. Kennedy who was assisted by Simulmatics (Lepore, 2020; Zuboff, 2019).

In summary, a company with the previously mentioned capabilities can, amongst other things, infer a person's sexual orientation, ethnicity, which religious beliefs they have, their political stance, personality traits, happiness and use of addictive substances (Kosinski et al., 2013), and is able to use this information for behavioural modification (Zuboff, 2019).

Since the misuse of AI can have such problematic consequences, it seems preferable that future development should not be primarily dictated by those who have the most to gain from exploiting unethical opportunities. Legislation can play an important role in putting the technology on the right path, but in order for it to truly benefit society as a whole, the democratisation of AI seems to be the path that needs to be explored.

1.3 The State Protecting the People

Nations have plenty of motivation to use AI as a technology. In many cases, the technology appears to be a clear benefit for the general public, such as integrating AI to assist in the urban planning processes (Wu and Silva, 2010), assisting doctors in the detection of cancer (O'Hare, 2017) and detecting water lines containing lead (Chui et al., 2018). Governments also play a very important role in regulating its use and protecting people's privacy through regulation. The General Data Protection Regulation (GDPR) seems to be one of the most discussed privacy regulation changes in recent years. It calls for changes in how consent to data collection and processing is collected, how cookies can be used and how a focus on data minimisation can be increased. The regulation also directly impacts the use of AI, stating in recital 71 that 'In any case, such processing should be subject to suitable safeguards, which should include ... to obtain an explanation of the decision reached after such assessment' (European Parliament, 2016). This is particularly relevant for machine learning, where one may not have a simple way of deciding exactly how the algorithm arrived at a decision.

There are other times at which the use of the technology is problematic, for example, how China uses facial recognition to recognise known criminals, suspects and jaywalkers (Mozur, 2018). It was recently revealed that Huawei was testing AI-based software with the goal of detecting people with Uighur characteristics – an oppressed minority group persecuted by the Chinese government – and triggering an alarm, potentially notifying the police about them (Harwell and Dou, 2020). Several countries, including the United States and the United Kingdom, use predictive policing to detect areas with high potential for crime and individuals who are likely to commit a crime (Couchman and Lemos, 2019; Friend, 2013), although police departments in the United States are backing off from the technology after the Black Lives Matter Protests (Lepore, 2020).

1.4 Shifting the balance of power

We are living in an information society, where the control of information and power, in many cases, can be considered synonymous. The act of possessing information is not in itself empowering; the power lies in the ability to use and control access to this information effectively and with credibility (Keohane and Nye Jr, 1998).

Before the internet, newspapers were an essential source of information for the general public. The journalists and editors of these newspapers became the guardians for the general public, seeing it as their responsibility to provide truthful, correct and objective information (Ward, 2009).

Today, the source of information has, for a lot of people, shifted from the newspaper and television to online news-feeds, search results and social media. There is, however, a profound difference in the structure of these new information sources compared to the traditional ones (Allcott and Gentzkow, 2017).

In some cases, such as the Facebook news feed, a single organisation controls the flow of information from authors and news outlets across the globe. Others, like Google Search, offer the ability to search the internet for information. However, this makes the Google Search service a single point of control for filtering and prioritising a vast network of sources, done automatically and autonomously. Both of these examples show companies in a position to affect people's view of the world by choosing what information to show and what to hide.

Companies behind such systems may have an interest in providing biased information, e.g., in order to increase sales or clicks. However, even when assuming that the intention is good, the technology in these systems will still not provide a neutral view of the world. The root cause of this is algorithmic bias, which can be explained as the tendency of algorithms to systematically behave in a certain way as a result of technical bias (technical constraints), preexisting bias (reflecting existing social attitudes) or emergent bias (differences between the context of development and use) (Friedman and Nissenbaum, 1996). Algorithmic bias also occurs in systems based on machine learning. The source of bias, in this case, is not the algorithm itself, but the dataset from which it learned its behaviour (Hajian et al., 2016).

The difference in power between the entity controlling AI and the entity in its scope is significant. AI can extract a lot of information about its subject, potentially without the knowledge of the entity. Whereas, the actor controlling the AI can use this extracted information for their own agenda. Therefore, when people in the general public are regularly in the scope of AI controlled by corporations, the balance of power is vastly out of balance. However, if we enable people with access to AI software and datasets, they may use this to turn the situation on its head. People can use the technology to analyse the behaviour of companies, extracting information regarding how their data is being processed, used and analysed, or they may analyse other information that may interest them.

1.5 Promote development of AI solutions that benefit the general public

Many of the AI technologies being developed at the moment are made to saturate corporate information needs. Thus, there may be an untapped potential for AI solutions that is useful for the general public, but is not of economic interest to corporations. Providing people with the data and tools they need to develop their own solutions may encourage the development of AI that taps into this potential. Further, by democratising the development processes used by companies, the general public can also influence the decisions made, such that the resulting AI is more aligned with the interests of the community.

I argue that the democratisation of AI is beneficial for the general public in three primary aspects.

1. **How AI is developed.** The development of AI can benefit from a wider community through the exchange of knowledge and expertise. Democratising the development would expand the current perspective from the view of a small group of experts to a complex community of people with different backgrounds and domain knowledge.

2. **How AI is used.** By democratising the use of AI, people would have access to the technology and could use it for the public good, developing applications that are aligned with public interests.

3. **How AI is governed.** Having AI managed by the community changes how the use and development of AI is evaluated, decentralising the discussion and enabling a greater spectrum of people to take a stance on what are acceptable norms and expectations beyond the legislation put in place by governments.

There are already a number of initiatives with the goal of making AI more beneficial to society as a whole, which is a step in the right direction, but they fall somewhat short of democratising AI, as they do not provide true democratic influence, do not even out the existing power imbalance, and they often are not open for the general public. For example, The Norwegian Data Protection Agency (Datatilsynet) has created the Sandbox for Responsible AI[1] where companies can get advice for developing AI that respects privacy. Further, the United Nations (UN) created AI for Good[2], a non-profit organisation with the goal of applying AI to address the UN's Sustainable Development Goals. Corporate AI does not appear willing to surrender actual power over AI, but, rather, is interested in contributions to technology development and identifying new use cases. For example, Microsoft's AI for Good[3] program provides organisations with access to AI tools developed by the company; however, they can also withdraw this access.

[1]Sandbox for Responsible AI https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/

[2]AI for Good https://ai4good.org/

[3]Microsoft AI for Good https://www.microsoft.com/en-us/ai/ai-for-good

The EU's Horizon Europe work programme draft for 2021 - 2022 highlights this exact issue. Under cluster 2, there is an upcoming call for research and innovation project proposals with the title *Artificial intelligence, big data and democracy*. The expected outcomes should include the following.

- Protecting fundamental rights and European values.
- Using AI to reinforce fundamental rights and European values.
- Introducing values-based frameworks for data governance and regulation of AI.
- Enhancing citizen engagement and democracy through the use of AI.

In order to further the development and use of AI in line with relevant rights and values, there is a need for a foundation that can inform the challenges that such democratic development faces, and the potential solutions and approaches that are proposed to address those challenges. As part of this thesis, I will propose a framework for informing such development, using the core ideas and principles of democracy as an approach by which to establish what the relevant rights and values are, how they may be challenged, and by which to decide how to choose between opposing sets of interests in the context of AI projects and platforms. The goal of the framework is to serve as a foundation for the future development of AI projects and platforms, informing choices during development regarding what technical components are integrated as part of the platform.

1.6 Research objectives

Establishing a framework requires a solid theoretical foundation, and, after going through the current literature on the topic, I found that there were no adequate overviews of the different aspects of democratising AI, nor was I satisfied by the definitions available. Thus, I have decided to create an overview of the literature on the democratisation of AI through a structured literature review. Further, I will use the overview to establish a more comprehensive definition, as well as create an overview of the various challenges and solutions for the democratisation of AI. Then, I will use the insights gathered through the previous steps to establish a socio-technical framework. Finally, I will conclude my thesis by reflecting upon my contributions and possible impact.

Here are the various goals of the thesis, summarised as a set of research objectives:

1. **Create a review of the literature on the topic** I will perform a structured literature review to create an overview of literature on the topic and establish a foundation to tackle my other research objectives.
2. **Establish a unified definition of the democratisation of AI** Using the existing literature, I will analyse how scholars define the term,

either explicitly or implicitly, and analyse how democracy as a fundamental concept can be related to AI, as well as identify and analyse central principles of the democratisation of AI.

3. **Establish an overview of challenges and solutions to the democratisation of AI identified by current literature** I will analyse the challenges identified by the current literature, as well as evaluate and summarise solutions proposed in the literature to these challenges.

4. **Develop a socio-technical framework for the democratisation of AI** I will use the insights gathered through the previous steps to establish a socio-technical framework that can be implemented in AI projects, as a platform or an infrastructure to further the democratisation of the development, use and governance of AI.

1.7 Structure of the thesis

My thesis is composed of six themed chapters. The first chapter, the introduction, provides an overview of problems relevant to the democratisation of AI, and of the motivation as to why it is an important and timely topic of research; also included are the research objectives I address throughout the thesis. In chapter two, the methodology, I describe and explain my choice of the structured literature review as my methodology for establishing a scientific and theoretical foundation for the rest of my thesis, as well as the limitations of my work. Thus, chapter two will address my second research objective. Chapter three addresses research objective two by analysing existing work on defining the democratisation of AI, establishing a relationship between the process of democratising AI and an existing definition of democracy, and drawing upon concepts and topics found in the literature review. The fourth chapter establishes an overview of the various challenges and solutions faced by the democratisation of AI, thereby addressing research question three. In chapter five, I address the final research question by drawing upon the insights gathered throughout the previous chapters to establish a socio-technical framework and propose a few ways in which it can be applied[4]. Finally, in chapter six, I will provide an overview of the work done throughout the thesis, including limitations of my work, how my work has contributed to existing research on the topic, and I will highlight future avenues of research.

[4]Appendix B expands upon this with a suggestion for how to implement a version of the framework as a technical platform for developing AI to evaluate privacy agreements.

2 Methodology

2.1 Context

To establish a theoretical and scientific foundation for the rest of my thesis, I will perform a structured configurative (Gough et al., 2012) literature review, searching for concepts and assumptions that are relevant for defining and analysing the democratisation of AI. The goal with this systematic literature review is to address my first research objective, *Create a review of the literature on the democratisation of AI*. I will use this literature review as a foundation upon which to establish a unified definition, in the next chapter, that builds upon and extends the various definitions used in literature. Further, I will summarise and analyse the various challenges and solutions found in the literature review in the subsequent chapter. Doing a structured literature review seems like the right approach, as it enables me to get an overview of what has previously been written on the topic, and thus provides me with a solid scientific foundation upon which to base the rest of my work. In addition, to the best of my knowledge, there does not yet exist any published literature review on the democratisation of AI. Therefore, producing a literature review using a structured and repeatable method provides new insights into an evolving field of knowledge, and seems like a valuable contribution in and of itself.

2.2 Limitations

I have constrained the search to primarily capture literature using some form of the term *democratisation* and that writes about AI, ML or data science. While there would probably be other valuable literature on the topic of the general democratisation of technology, I have chosen to specifically focus the search terms on AI, in order to keep the scope down, as there is a time constraint on this thesis. Given more time, I would delve deeper into each topic revealed in the review and explore how democratisation of other areas could inform this topic. Further, other terms such as *democratisation of big data*, *open AI*, *accessible AI* and *social AI* could potentially uncover more related literature and might be worth exploring in future work.

While I will be covering all topics that arises in the literature, I will not delve in depth into them all, as the fields cover everything from politics to AI and data science to participatory design to international relations. Delving into all of these topics would require a great deal more time and a much deeper understanding in many of the fields than my computer science background enables, so my intention is to cover the arguments from the literature review as an overview and instead encourage further research into each topic by researchers in the respective fields.

2.3 Structured literature review

My systematic literature review methodology is based on the processes described by Bryman (2016), O'Leary (2017) and Seale (2018). As the goal of the review was to create an overview of the current discussions about the democratisation of AI, as well as the definitions and perspectives on what

the term means, I decided to focus on capturing a wide variety of perspectives rather than narrowing the results down to the papers of the highest possible quality, which would have been prudent for other goals. In the process, I still filtered out papers that I considered to be of too low quality, but I would have been even more restrictive if the review had been for another purpose, such as a meta-analysis summarising research findings.

2.3.1 Search

For this structured literature review, I chose to utilise the search engines Google Scholar, Scopus and Web of Science, as they were recommended by the university library. Rather than performing a separate scoping review to determine what search terms I would use, I instead applied an exploratory approach wherein I started out with a number of intuitive search terms, such as *democratisation of artificial intelligence* and *democratised AI*, and expanded the queries with relevant terms mentioned in the resulting texts. Most notably, the term *artificial intelligence* was expanded with relevant terms such as *machine learning*, *deep learning* and *data science*.

The term *beneficial artificial intelligence* was also explored, but it had a tendency to return results regarding artificial general intelligence and how to ensure such a technology would behave ethically. Although the results in Scopus and Web of Science were usable, Google Scholar returned results that diverged greatly from the topic in question, and I aborted the evaluation after evaluating 400 search results. I also considered *responsible artificial intelligence* which returns results regarding policies, ethics for AI development and use, and ethically reasoning AI agents. These topics seemed quite relevant for some aspects of the democratisation of AI, barring the articles about ethically reasoning AI agents. However, the topics were also surfaced by previous queries, and the time constraint of the thesis limits the scope of my search; thus, I chose not to include that query.

The terms *democratisation* and *democratised* can be spelled either with an 's' or a 'z', depending on whether the author is writing in U.S./international English or British English. Different results were returned by the search engines depending on what spelling was used, so I ensured that the queries covered both possibilities. It appears that the U.S./international spelling is more popular in the current literature, based on the number of results for each query (see table 1).

The use of * as wildcard, as well as *AND* and *OR* for logic queries, were very practical in both Scopus and Web of Science for expanding the scopes of the queries. Google Scholar, however, does not appear to properly support these functions. Hence, I restricted the use of Google Scholar to search for specific phrases, such as *democratised AI* and *beneficial AI*.

In the beginning of the search, I tried to avoid creating queries that were too broad, in the fear of large amounts of irrelevant results, but I expanded the query scope as I realised how few articles were published on the topic. This caused some of the articles to be returned by multiple queries. In other words, the search results listed in the Table 1 do not represent unique articles across all queries, but include articles that were surfaced multiple times. However, as I removed duplicate papers in the screening of search results, the numbers in the columns *Pass 1* and *Pass 2* are counts of unique

Table 1: Literature review search queries

Search engine	Query	Results	Evaluated	Pass 1	Pass 2
Google Scholar	”democratization of artificial intelligence”	83	83	24	7
Google Scholar	”democratization of ai”	153	153	55	15
Google Scholar	”democratized artificial intelligence”	36	36	8	4
Google Scholar	”democratized ai”	25	25	4	1
Google Scholar	”democratisation of artificial intelligence”	4	4	1	0
Google Scholar	”democratisation of ai”	24	24	2	1
Google Scholar	”democratised artificial intelligence”	0	0	0	0
Google Scholar	”democratised ai”	4	4	0	0
Scopus	TITLE-ABS-KEY (”democratization of artificial intelligence”)	9	9	0	0
Scopus	TITLE-ABS-KEY (”democratisation of artificial intelligence”)	9	9	0	0
Scopus	TITLE-ABS-KEY (”democratization of AI”)	4	4	0	0
Scopus	TITLE-ABS-KEY (”democratisation of AI”)	4	4	0	0
Scopus	TITLE-ABS-KEY (”democratized artificial intelligence”)	2	2	0	0
Scopus	TITLE-ABS-KEY (”democratised artificial intelligence”)	0	0	0	0
Scopus	TITLE-ABS-KEY (”democratized AI”)	0	0	0	0
Scopus	TITLE-ABS-KEY (”democratised AI”)	0	0	0	0
Scopus	(TITLE-ABS-KEY (democratization) AND TITLE-ABS-KEY (artificial AND intelligence))	63	63	11	1
Scopus	(TITLE-ABS-KEY (democratisation) AND TITLE-ABS-KEY (artificial AND intelligence))	63	63	0	0
Scopus	(TITLE-ABS-KEY (democratization) AND TITLE-ABS-KEY (ai))	28	28	2	0
Scopus	(TITLE-ABS-KEY (democratisation) AND TITLE-ABS-KEY (ai))	28	28	0	0
Web of Science	ALL=(”democrati*ation of artificial intelligence”)	4	4	0	0
Web of Science	ALL=(”democrati*ation of ai”)	2	2	0	0
Web of Science	ALL=(democrati*ation AND ”artificial intelligence”)	36	36	2	1
Web of Science	ALL=(democrati*ation AND ai)	31	31	0	0
Scopus	TITLE-ABS-KEY (democra* AND (”artificial intelligence” OR ai))	534	534	39	17
Web of Science	ALL FIELDS: (democra* AND (”artificial intelligence” OR ai))	618	618	11	4
Scopus	TITLE-ABS-KEY (democra* AND (”deep learning” OR ”machine learning”))	285	285	9	8
Scopus	TITLE-ABS-KEY (democra* AND (”data science”))	54	54	9	4
Web of Science	ALL FIELDS: (democra* AND (”deep learning” OR ”machine learning” OR ”data science”))	261	261	5	3
Scopus	TITLE-ABS-KEY (”beneficial artificial intelligence”)	4	4	3	2
Scopus	TITLE-ABS-KEY (”beneficial ai”)	9	9	6	1
Web of Science	ALL FIELDS: (”beneficial artificial intelligence” OR ”beneficial ai”)	11	11	2	0
Google Scholar	”beneficial artificial intelligence”	774	400	11	7
Sum		**3162**	**2788**	**204**	**77**

papers, where papers are counted in the first query in which they appear.

2.3.2 Screening search results

To decide what papers are relevant and which to discard, I needed some definition of that for which I am looking. In this case, however, since I am trying to construct a definition based on the literature review, it is important to avoid settling on a definition while doing the screening. Otherwise, I would simply end up eliminating the papers that do not fit the original definition, which would become a self-fulfilling prophecy. To keep this from happening, I decided to start with a very wide scope, keeping any papers addressing democratisation and AI.

As part of the exploratory process, I defined a set of topics that I considered related to the democratisation of AI. The set of topics started out quite narrow, but, as I processed the search results, I added new topics that I came across to the set, being careful to err on the side of capturing too many topics rather than than too few. The final set of related topics is listed below in the set of conditions for eliminating search results.

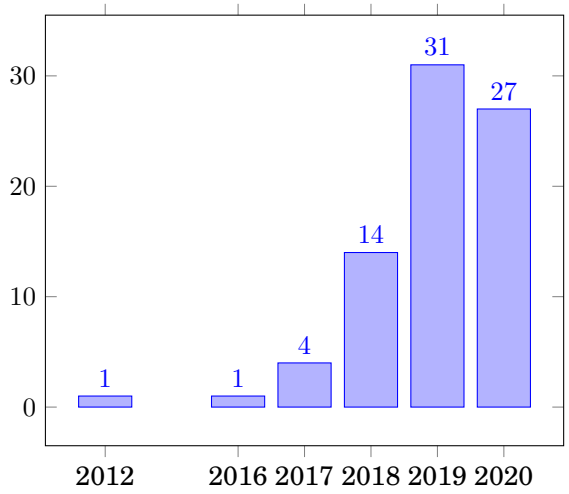
In order to decide what papers I would keep and which to discard due to irrelevancy, I devised a set of criteria defining the boundaries of what I consider relevant. Any paper meeting one or more of the following criteria was excluded from the literature review:

- only mentions AI and democratisation in separate contexts;
- uses AI and democratisation in the same context, but only writes about using AI as a tool to democratise a different activity;
- merely mentions the democratisation of AI, but without any real substance;
- does not write about democratisation, governance, policies, norms, ethics, social impact, human rights, transparency, inclusiveness, diversification, participatory design or development, globalisation, balance of power, fairness, bias, crowd sourcing, data exploration, data labelling, data cleaning, data processing, data sharing, software sharing, knowledge sharing, education, simplified development, integration, auto ML, explainable AI or validation in some way relatable to AI.

In the screening of search results, I went through all those returned for each query. There was one exception, however, which was the final Google Scholar search on beneficial artificial intelligence. For that query, I processed the first 400 results, but only the first few result pages were at all relevant to the democratisation of AI. Thus, I decided not to process the remaining 374 results.

There is a notable difference between how the three search engines display their search results. Scopus and Web of Science have the article abstract available in the search results, but Google Scholar displays sections of the text that are predicted to be relevant to the query or that contain the search terms used. Therefore, I applied two different screening strategies, one for Google Scholar and one for Scopus and Web of Science.

Figure 1: Papers selected in the literature review, counted by year of publication



For Google Scholar, if the title was clearly irrelevant, the highlighted sections appeared irrelevant and the search term was either not highlighted or appeared exclusively in the bibliography, in which case I discarded the result.

For Scopus and Web of Science, if the title and abstract were clearly irrelevant, I discarded the search result.

2.3.3 Screening articles

To screen the articles, I checked to see if any section of the article seemed relevant, essentially looking for the inverse of the irrelevance criteria from before. If I found a section that seemed relevant, I would check the quality of the paper. To judge the quality, I used Google Scholar to look up the H-index of the authors of the paper and the journal or conference series in which it was published, as well as to evaluate the structure and content of the paper itself. If the author and the conference or journal were unlisted or had a very low H-index in relation to others in the same field, I would be very reluctant to include the paper. I decided, however, to make an exception if the paper contributed an interesting perspective or insight, since the purpose was to create an overview of the various topics and approaches suggested. However, if the paper did not provide any real contribution towards the objective, was poorly written or otherwise seemed less than serious, it was discarded.

In Figure 1, I have counted how many papers were published in each year. The resulting Figure seems to indicate a growing popularity since the concept's inception in 2016[5], as more papers were published in each

[5]In the literature review there is a paper from 2012 by Buckingham Shum et al., about the democratisation of open data, complexity science and collective intelligence, mentioning AI as

consecutive year. In 2020, the number of papers published was somewhat lower than that of the previous year, but I attribute this feature to the fact that I performed the search in 2020, and thus there may have been other papers published later in 2020 that were not included in this count.

The fact that papers start appearing in 2016 is also noteworthy. One reason for this timing may have been that AI started becoming more useful for practical applications around this time. A notable breakthrough to illustrate this is a paper by He et al. (2015), in which they demonstrate a neural network exceeding human level performance in the task of recognising objects in images (He et al., 2015). This transformation in the conceptualisation of AI, from a technical curiosity into a practical tool that may solve real world problems, seems like a possible explanation for the newfound interest in democratising the technology.

2.3.4 Exploratory categorisation

To identify what topics are included in the literature in the definition of the democratisation of AI, I performed an exploratory categorisation of the topics related to the term in the papers. These categories include meta topics, such as explicit discussion about the definition of the term, the current status of democratisation and challenges that democratisation must tackle, as well as a topic for each challenge identified where there is a paper discussing an approach to the challenge.

Table 2 provides an overview of the various categories I identified, noted in the first column. Further, in the second column I noted the sub-topics I found in the literature related to each category. In the third column I have included the amount of papers that fell into the related category. As I decided to include papers that have a section discussing one of the related topics in a category, rather than only count papers dedicated to the topic, I have included a number of papers that are represented in multiple categories.

While analysing the various articles, I identified another, more refined, set of categories and topics in relation to the democratisation of AI. This new categorisation is founded upon the five principals of democratising AI, which will be discussed in the next chapter. Since the categories are focused on the democratisation of AI, they do not include the meta categories from the previous set.

Several of the primary categories were highlighted by scholars as principles of democratisation, for which I found support through an analysis of democracy. Further, the principles were central concepts relating to the topics discussed in the broader literature. However, some researchers have pointed out additional principles for which I did not find support, and thus chose to discard. There were also some definitions of the principles with which my analysis did not align, so I altered the definition of those principles to better fit my findings. In addition, I added a principle of my own,

part of this process. However, since the authors do not explicitly write about the democratisation of AI, I rather consider the paper as a broader precursor to the concept. In contrast, Nakamura and Yamakawa (2016) uses the term *democratisation of AI*, noting briefly that enabling more people to create their own AIs is a step in the process of democratising the technology.

Table 2: Literature review categories

Category	**Topics**	**Number of papers**
Definition	Definition	5
Current status	Existing solutions	3
Challenges	Corporate vs community Centralised vs decentralised Open vs closed development The technological race for AI Transparency Inclusiveness Convergence Global South	21
Knowledge development	Education Knowledge sharing Diversify field	11
Inclusion	Fairness Bias Participatory design and development Social impact Establishing norms	19
Governance	Fundamental rights Policies	12
Data access	Data sharing Crowd-sourced data labelling Data cleaning	11
Interpretability	Data exploration Explainable AI	16
Simplified development	Auto ML Validation	12
Software sharing	Software sharing	5

Table 3: Refined literature categories

Category	**Topics**
Decentralised control	Challenges with centralised control Interoperability Democratic governance Governing open resources
Accountability	Ethical principles Regulation Unemployment Approaches for accountability
Transparency	Algorithmic transparency General transparency
Openness	Dual-use Race for AI Access to data Data exploration Auto ML Access to hardware Data preprocessing Access to AI Education
Inclusiveness	Fairness Stakeholders Participation Communication Beneficial AI

accountability, which was not highlighted as a principle, but was a key concept in relation to the discussions regarding the regulation and governance of AI.

Within each category, I have listed a set of topics that were central discussions in the literature reviewed. Some of the topics are relevant for several of the categories, as the categories share a number of aspects. In those cases, I placed the topic in the category to which the topic's discussion contributed the most. An overview of the categories and topics is provided in Table 3. I decided to use this set of categories and topics to organise my chapter on the challenges and solutions of democratising AI.

3 Defining the democratisation of AI

The term *democratisation of AI* seems to have first appeared in a paper from 2016, by Nakamura and Yamakawa, titled *A Game-Engine-Based Learning Environment Framework for Artificial General Intelligence*. Since then, the number of papers discussing the topic and related subjects appears to have grown year over year, as shown in Figure 1.

However, while the existing literature offers two explicit definitions for the democratisation of artificial intelligence (namely Ienca, 2019; D. Wang et al., 2020), neither definition seems to capture the concept in its entirety. Looking at how the different papers use the term reveals differing understandings about what the term encompasses. Some authors (such as Masood and Hashmi, 2019) approach the topic as an entirely technical challenge, while others (see for instance Buckingham Shum et al., 2012; Ienca, 2019; Moreau et al., 2019; Sudmann, 2020) recognise that the topic also extends into politics and sociology. To address this issue, I will analyse how the existing literature writes about and uses the term, as well as relates the concept of democracy to AI. My purpose is to create a unified definition that captures the various insights revealed in the analysis, addressing my second research objective: *Establish a unified definition of the democratisation of AI*.

While the democratisation of AI is an emerging topic in scientific literature, it is important to realise that the term and the definitions do not exist in a vacuum. First and foremost, they draw on literature that builds upon the democratisation of technology, a topic that has been studied for decades (Feenberg, 1991). Further, they also rely on observations and understandings of democratisation of governance. In particular, implicit within the authority of the democratic state is the establishment and enforcement of laws and regulations necessary for ensuring responsibility and accountability of the different actors in the field, as well as the decentralisation of political control to ensure participation in various forms. Apparently, in many states, democratisation of AI faces great hurdles, due to states using undemocratic forms of power. In states where elections are not free and fair or where the government is no longer truly accountable to the population due to the amount of secrecy or external influence on the democratic processes, conditions are insufficient to regulate the development and use of AI in democratic ways. In this case, participatory and communicative discourses are recognised as alternative strategies by which to achieve democratic processes, implying both national and international cooperation, agreements and pressure. Thereby, people and advocacy groups may dissuade governments from abusing the technology to maintain undemocratic control, wherein they deploy the technology to silence opposition, impose surveillance or develop unethical AI systems such as Lethal Autonomous Weapons (LAW).

There is a gap between how companies through their websites and a majority of scholars describe the current status of democratisation. Big companies such as Google and Microsoft and a small number of scholars use the term democratised AI, referring to auto AI solutions (Sudmann, 2020)). This implies that the tech companies have reached a state wherein AI has

become democratised as an outcome and that the process, therefore, is in some way completed. On the other hand, the vast majority of the literature disagrees, noting current and anticipated issues and challenges with how AI is designed, used and governed (Ienca, 2019; Sudmann, 2020). One can interpret the narrative that the companies use as an attempt to gain good faith and public trust while they are simultaneously wielding this technology for their own economic benefit. They also refer to the term to highlight how technology is used for public good, such as how Microsoft participates in projects addressing climate change, while attempting to distract from negative cases, as in the case where Microsoft also cooperated with China's National University of Defence on AI problems that commentators believed could be used for state surveillance (Sudmann, 2020). It is, therefore, necessary to review the literature independently of such corporate interests.

3.1 Analysing the democratisation of AI as a term

We can glean a few insights regarding the democratisation of AI by merely analysing the term by itself. First of all, democratisation is a process (Ienca, 2019), implying that AI is inherently undemocratic (Sudmann, 2020). This point was argued in the introduction of this thesis, highlighting how the primary technical components of AI greatly shift the balance of power in favour of the parties making use of the technology. As such, the current use of AI, in many cases, is very problematic from a societal and democratic point of view.

Further, researchers considers the term to be of a utopian-idealistic nature (Sudmann, 2020). While the process itself seems within reach, the goal, like the goals of other types of democratisation, in reality appears to be facing practically insurmountable obstacles. For example, people should be equal in the development and use of AI (Ienca, 2019; Sudmann, 2020), but, in practice, a number of people will have impairments or simply not have access to computers or the internet, diminishing their ability to participate on an equal footing. This is not to say that one should not strive to reduce this disparity, however, as some level of equality is more desirable than none.

3.2 Defining democracy in relation to AI

Despite there being a few attempts at defining the democratisation of AI, none of the authors elaborates on how their works draw on some established definition of democracy. Thus, to address this apparent gap in the literature, I will first have to establish what democracy can mean in this context. David Held (2006) defines democracy as '(...) a form of government in which, in contradistinction to monarchies and aristocracies, the people rule. Democracy entails a political community in which there is some form of political equality among the people.' (Held, 2006, p. 1).

From Held's definition of democracy, there are two points that carry particular significance in relation to the democratisation of AI: *the people rule* and *political equality among people*. Similar to Held (2006), other authors who write about the democratisation of AI highlight equality and rule of the people as central concepts (see for instance Ienca, 2019; Sudmann, 2020).

I argue that both aspects are relevant in this thesis, which leads to the following questions:

1. What do the people rule?
2. Who are the people?
3. What is political equality in the context of AI?

3.2.1 What do the people rule?

In a traditional democratic context, the object in question would naturally be the democratic state, which means that the citizens with voting rights, in principle, rule the decisions regarding the governance of the democratic state. This can be achieved through representative systems, wherein the parliament legitimises politics and authorises governing bodies that manage the society according to the rule of law and by the use of expertise. In regard to the democratisation of technology, I suggest this is equivalent to people ruling the decisions regarding the management of AI, such that the management of AI is not limited to decisions regarding the governance of AI, but also includes decisions within the various processes in the development and use of AI.

3.2.2 Who are the people?

The question about who the people are, in relation to a democracy, dates back to ancient Greece and remains, to this day, an open question (Held, 2006). To help answer this question in the context of AI, there is an intermediary question that can be answered: who has a legitimate claim to exert influence over the outcome of a decision? A natural answer to this question is: anyone who would in some way be impacted by the outcome of the decision, whether directly or indirectly, has a legitimate claim to influence the decision in direct proportion to the decision's impact on them. In other words, everyone who is concerned, has an interest, or is affected, now or in the future as AI develops.

Further, the Legitimate Claim to Influence (LCI) seems to align with Hutt's (2018) principle of Equality of Access and Deliberation: 'EAD comprises two sub-principles: equality of access, and equality of deliberation. Equality of access means that all those potentially affected by collective decisions must have an equal opportunity of entering the fora where those decisions are adopted. Equality of deliberation requires that decision-making processes be sensitive enough to be able to capture, make visible, and consider the claims of all the participants in the discussion in a non-dominating manner.' (Hutt, 2018, p. 98). A notable difference between EAD and LCI is that LCI grants differing level of influence depending on the impact on the participant, whereas EAD leaves this decision to the parties involved in the deliberation process.

Moreover, the legitimate claim to influence in the context of a deliberative process can be transferable, meaning that a person or organisation who represents someone else that has an LCI can wield this claim to further the interests of the person they represent. This distinction enables, for

example, advocacy groups to defend the interests of groups who may not be able to defend their interests on their own.

There are some practical issues relating to LCI. For example, determining who actually has an LCI becomes an undecidable problem, since it is impossible to determine the exact future implications of a decision. However, one can address this issue by, for example, limiting influence to those who face a significant foreseeable impact from the decision. On this backdrop, the literature identifies a number of stakeholders relevant for the process of democratising AI:

- **AI researchers** include experts from academia, public institutions as well as companies with a focus on developing the state of the art in AI (Moreau et al., 2019). These experts stay up-to-date and contribute to the state-of-the-art in the field (Moreau et al., 2019).
- **AI developers** may contribute to innovation in a particular topic or application but primarily use the technology for their own purposes, without necessarily staying up-to-date with the state-of-the-art (Moreau et al., 2019). This category includes Small and Medium-sized Enterprises (SMEs) that employ AI in their work, but without specialising in the technology (Moreau et al., 2019).
- **AI users** use ready-made AI models to address needs they may have, such as, for example, as part of an app or website.
- **Politicians** (Buckingham Shum et al., 2012) participate in the development of new regulation and legislation regarding the design, development and use of AI.
- **Advocacy groups** (Buckingham Shum et al., 2012) work to protect the interests of the groups they represent.
- **General public** is made up of everyone else, those individuals who do not have any direct connection to or significant knowledge about AI, but may still use it as a part of end-user applications (Moreau et al., 2019) or who are subject to it as part of their daily lives.

It is important to note that there are majority and minority groups within the various stakeholder groups, in particular the general public, as relate to historic or continued oppression, disparities in access to resources, education, and so forth.

I will use the term non-experts to refer to AI users, politicians, advocacy groups and the general public, while excluding people who have adequate knowledge and experience in AI to develop AI for their own needs without the use of greatly simplified development processes.

3.2.3 What is political equality in the context of AI?

There is a variety of different types of democracies, each of which interpret equality in a different way. The different types seem to fall within three different categories, depending on exactly the aspects with which they are concerned, in regards to equality. Further, the various categories can also

help clarify what types of democracies are compatible and which are contradictory, as each type of democracy seems to be less compatible with the others within the category. However, one may pick and choose different types of democracies across the categories in order to assemble interpretations of equality that capture different considerations than would any distinct type by itself.

The first category concerns themselves with whether or not it is fair to impose restrictions in order to distribute resources more evenly throughout the democracy. In this category we find the liberal democracy, which considers not only that citizens should have equal rights, but that there should be a focus on the freedom of the individual rather than equality among peers in other aspects of their lives (Beretta et al., 2019). We also find the egalitarian democracy, which aims to ensure complete equality, meaning that participants should not only have equal rights, but equal opportunities (Beretta et al., 2019).

The second category is instead concerned with how decisions can be reached. Here we find the deliberative democracy, in which the focus is on enabling a discussion between peers to arrive at rational solutions and compromises, where everyone should have equal rights to have their arguments and voices heard as well as taken into consideration (Beretta et al., 2019; Elster, 1998). However, equality in the process of deliberation has further implications, as it gives people the right to a minimum of living standards, education and other benefits (Hutt, 2018). The purpose of this right is to ensure that people will not be dominated in the process of deliberation by others with, for example, a better socio-economic background (Hutt, 2018). Further, equality in deliberation also emphasises that people without deep technical knowledge should have an equal opportunity to participate, rather than limiting deliberation to groups of experts (Habermas, 2015). In contrast, a competitive democracy means ensuring that participants have equal opportunity in asserting their positions and decisions, with the compromise that someone else in the future will have the opportunity to assert their own decisions (Beretta et al., 2019).

Finally, the third category consists of a solitary type of democracy, the republican democracy. A republican democracy aims to hold leaders accountable to the public for their choices, and thus focuses more on enabling equal rights among citizens to challenge decisions made by the leaders, than arriving at decisions through public deliberation (Beretta et al., 2019). The goal of this focus on accountability is preventing people from being subject to domination by others. One can argue that the republican democracy is in somewhat of an opposition to a deliberative democracy, but I will suggest that the two exists more as trade-offs with one another, rather than in actual fundamental disagreement. For example, any deliberative representative democracy relies on a balance between the republican and deliberative views of equality.

In this thesis, I will borrow central dimensions from the deliberative, egalitarian and republican perspectives in my interpretation of equality, tackling challenges and solutions that not only concern themselves with fundamental rights, but also aim to ensure accountability of the various actors and to minimise disparities related to AI, such as access to AI resources and knowledge, economic opportunity and influence on the decisions re-

garding AI. The reason for this is that this interpretation of equality seems to be the most ambitious of the various options, and thus for the other options some challenges or solutions may not be relevant. For example, in a purely liberal interpretation of equality, one does not need to consider that certain groups, such as large companies, may have better access to AI technology than others, as it does not necessarily infringe on any rights. By contrast, from an egalitarian perspective, the groups with access have a better opportunity to benefit from the technology than do the rest, and thus one should attempt to increase the general public's access to the technology.

3.3 Principles of democratisation

I would like to highlight that, while the notion of principles can seem quite passive, as they assert what is or what should be, in this case each principle can also be considered a mechanism by which to further the process of democratising AI. This small shift in perspective can help illustrate how the principles are not just passive descriptors, but also serve the definition as an active component of the process.

In the following, I will be considering different dimensions of democratisation of AI in order to synthesise a more coherent set of concepts, processes and models. The resulting collection of principles is sufficient to capture all challenges and solutions from the literature that seemed relevant. However, the relationships between the principles are somewhat complex, as the realisation of each principle depends on certain aspects of the other principles, and thus some topics in the democratisation of AI are related to several principles.

To illustrate how concepts were selected, evaluated and organised, I provide this example. The principle of interoperability is mentioned by Ienca (2019) in their definition of the principle convergence. The next question is, then, do any of the existing principles already supersede interoperability? On the one hand, one can consider interoperability as a mechanism by which to prevent AI developers from being locked into a single ecosystem, as their artifacts would be compatible with other frameworks. From this perspective, it is natural to argue that interoperability belongs within decentralised control. On the other hand, interoperability is also a way of keeping the technology open, as it relies on at least partially open standards in order to enable parsing by, or communication with, other frameworks. Thus, one can also argue that interoperability belongs within openness. However, I will suggest that interoperability contributes more to the decentralisation of control than it does to the openness of AI, and I have therefore chosen to categorise this principle within decentralised control.

3.3.1 Decentralised control

First of all, the concept of decentralised control seems to be one of the most fundamental aspects of democracy, as it directly relates to the rule of the people. Thus, the control of AI should be decentralised to avoid accumulation of economic, political and technological power in the hands of a single or a few cooperating actors (Buckingham Shum et al., 2012; Ienca, 2019).

This necessity holds true, not only for corporate, but also for state or international entities, as they may hold a disproportionate amount of power over other people, companies or states (Ienca, 2019).

Ienca (2019) defines convergence as the interoperability, intercommunication and ease of integration as well as the convergence of neurotechnology and AI, and argues it is a principle of the democratisation of cognitive technology. The convergence of neurotechnology and AI is not relevant to my work, as I focus solely on AI. However, the rest of the definition seems to be covered by my definition of decentralised control, as interoperability, intercommunication and ease of integration are all mechanisms of decentralising control. Thus, I have chosen not to include convergence as a separate concept.

3.3.2 Accountability

Accountability in the democratisation of AI means ensuring that actors developing, using and governing AI and related resources can be held accountable by the stakeholders of their AI and decisions, similarly to how democratic governments are to be held accountable by their citizens. This is a necessity for balancing the power between those who use AI and those who are subjected to AI, as well as for maintaining trust in how the technology is being used. Accountability relies on decentralised control, transparency, openness and inclusiveness, as without decentralised control the accountability may not be democratic, without transparency it may be harder to determine whether the AI is being misused, without openness it becomes harder to explain what causes the issue or there may be a lack of knowledge necessary to make good decisions regarding the technology, and without inclusiveness the impact on certain minorities may not get adequate attention.

3.3.3 Transparency

Transparency enables insight into the processes, decisions and results from the design, development, use and governance of AI. This may include information about which stakeholder groups were involved in reaching certain decisions, what steps were taken to mitigate negative bias in the final AI model, what data is used by an AI system, how reliable the system is or information about how a system reaches a conclusion. Transparency is also related to the rule of the people, as a necessity for ensuring that accountability is enforced.

One may argue that transparency around the development, use and governance of AI is a necessity for ensuring the entities controlling AI can be held accountable for their actions. As well as enabling trust between those who control AI and the general public (Buckingham Shum et al., 2012), transparency is a way of balancing knowledge and power (Belinchon et al., 2019).

Ienca defines transparency as 'the principle of enabling a general public understanding of the internal processes of cognitive technologies[6]' (Ienca,

[6]Ienca (2019) defines cognitive technology to include AI earlier in the paper.

2019, p. 275), however, this definition seems too narrow. While insight into the internal AI processes may lead to increased awareness of how an AI system reaches a decision and the ways in which the decision may be incorrect, the definition does not include information regarding the planning, development, use or governance of AI. Instead, the definition seems more akin to the term algorithmic transparency, which is discussed in section 4.3.1. However, Ienca (2019) goes on to acknowledge the need for transparency regarding when and why AI is applied, the sources of the data and expertise that were used, as well as data protection and ownership, even if it does not line up with their definition of transparency.

3.3.4 Openness

Openness regarding AI means there should be fair access to the technology (Ienca, 2019) and the resources necessary for its development and use, as without access one cannot participate in the development and use on an equal level with others. One can certainly argue that the concepts of transparency and inclusiveness naturally belong within the concept of openness, but I have purposefully separated the three concepts for clarity, because they are relevant as their own separate discourses in and of themselves.

Ienca states that 'openness in [cognitive technologies] involves the principle of infusing every application we interact with, on any device, at any point in time, with (components of) cognitive technology' (Ienca, 2019, p. 275) and argues that it would 'prevent their uneven accumulation among restricted applications or tools' (2019, p. 275), which seem not to be an obvious implication of the term. While the availability of AI technology for the applications where it would be useful seems to be a sensible interpretation of openness, the necessity of applying the technology in all applications seems to be beside the point and most likely impractical. The accumulation of AI technology in applications where it is useful, and the absence of the technology in applications where there are other, more practical, solutions, does not seem to pose any threat to the democratisation of AI, but rather appears to be a pragmatic approach to application development.

3.3.5 Inclusiveness

Inclusiveness in regard to AI means that there should be equality in the development, use and governance of AI among all stakeholders; no social group should be unfairly treated. This also implies that non-experts should be included in the development, use and governance of AI, relating back to the discussion on equality.

Some authors use terms such as marginalisation and being left behind (see for instance Ienca, 2019), but I find that these words create a narrative of a necessity for everyone to make use of AI. In reality, some social groups might prefer to avoid AI, whether due to religious, cultural or other interests. Instead, the word fair is context-dependent and reflects the concept that everyone should have the opportunity to make use of and benefit from AI, yet they should also have the option not to be subject to it.

Further, the definition of inclusiveness also reflects that there are degrees of fairness and that, in many cases, there is no option that is equally

fair for everyone. Therefore, stating that *no one should be treated unfairly* rather than that *everyone should be treated fairly* emphasises that the goal is to avoid situations in which one or several groups are discriminated against for the benefit of others, even if the discriminating solution may be best for the average person in the community.

Sudmann (2020) identifies social good as a principle of democratisation and argues that the democratisation of AI implies that the technology should be used for social good. While I agree that the democratisation of AI is related to social good, I am inclined to disagree that it is a necessary principle. I will instead argue that democracy is not inherently about what is good and what is bad, but is instead about equality and fairness. My premise is the assumption that a majority of people wants what is best for society. Thus, the relationship between social good and democratisation is rather a consequence of decentralised control and inclusion, and, as with the influence on a decision, people can implement their own perspectives on what is good and fair. In their article, Sudmann (2020) uses the concept of participation rather than inclusion, which may help explain why they found the need for including social good as a separate concept. Participation seems to be a narrower concept, which does not necessarily consider the implications for those not actively participating. Inclusiveness, on the other hand, covers both the ability to participate and the consideration of groups who may not be able to participate.

Finally, Ienca (2019) distinguishes user-centredness as its own principle for the democratisation of AI. Nevertheless, I will argue that it seems to be more of a design principle than necessarily an aspect of the democratisation of AI, and that the benefits it might bring to the process would be covered by the inclusiveness concept.

3.3.6 The layers of the democratisation of AI

Based on the principles above, we can conclude that the process of democratising AI happens in three layers: technical, social and political. The technical layer relates to technical aspects of AI, such as whether it is accessible and transparent. Further, the social layer relates to social aspects surrounding AI, for example who has access to the technology and how choices during the design of AI impact stakeholders. Finally, the political layer relates to the control of AI and the related processes, whether through regulation or informal processes. These three layers do not exist in isolation, however, as the layers can inform decisions or provide solutions to problems discovered within other layers. Djeffal (2020) also writes about the technical, social and governance aspects of AI, but focuses on understanding the choices made about AI rather than the process of democratising AI.

3.4 Previous definitions

In the literature review, I came across two explicit definitions related to the democratisation of AI. The first definition is by D. Wang et al. (2020): 'Researchers coined it as "democratizing AI", where non-technical users are empowered by AutoAI technologies to create and adopt AI models.' (D. Wang et al., 2020, p. 77)

The second definition is by Ienca (2019): 'Consequently, democratizing cognitive technology implies a process of decision making about CT that will guarantee a possibility of fair access to CT for all users and a principle of equality among users during various stages of decision-making (including design, development and application)' (Ienca, 2019, p. 272)

D. Wang et al. (2020) appears to understand the term as a result of Auto ML, which seems to be a very narrow understanding of the term. Whereas, Ienca (2019) has a broader perspective, highlighting that it is about equality in the decision making about the technology, and that fair access to the technology is also relevant. However, there are a number of different aspects identified above, that neither definition address. For example, Ienca focuses on equality among users, whereas I argue that there are other stakeholders who should also be included in this equality, such as those who do not use the AI but are subject to it.

3.5 Definition

One of my research objectives was to establish a unified definition of the democratisation of AI. By compiling the various insights acquired through analysis of the term *democratisation of AI*, analysis of how democracy as a fundamental concept can be related to AI, as well as the identification and analysis of five principles for the process of making AI more democratic, I have arrived at the following definition of the democratisation of AI.

> The democratisation of artificial intelligence is the technical, social and political process towards accomplishing equality among stakeholders in the development, use and governance of AI through decentralised control, accountability, openness, transparency and inclusiveness.

This definition will inform decisions made in the next chapter, relating to what challenges and solutions were selected from the literature, and it will provide a structure, based on the five principles of the democratisation of AI, from which to organise the various challenges and solutions in the subsequent chapter.

4 Challenges and solutions to the democratisation of AI

In this chapter, I will go through the different challenges and solutions I identified through the literature search. The purpose of this chapter is to tackle research objective number three: *Establish an overview of challenges and solutions to the democratisation of AI identified by current literature*. Further, this chapter will serve as a foundation for the framework being established in the next chapter by identifying the various challenges the framework can address and a number of possible solutions that can be used to address them.

This chapter builds upon the definition established in the previous chapter to evaluate, structure and organise the various challenges and solutions from the literature. Thus, the challenges and solutions have been separated into five sections, one for each principle in the definition, to provide structure. However, some challenges and solutions may fit several sections due to how there is a certain overlap between the principles and how solutions may have more than one aspect to them. In those cases I have evaluated what principles the challenges or solutions contribute the most to, and placed them where I think they are most relevant, even if one certainly could argue for placing them somewhere else.

To begin with, I will present methods of evaluating how democratic the processes a project uses for development and use of AI is, as this topic necessarily encompasses all the concepts from the definition. In the literature I found a number of frameworks proposed for evaluating these processes, and while the frameworks may not be specifically about democratisation or cover all the challenges of democratisation under my definition, they may still serve as good starting points for guidance and evaluation.

Garvey (2018) suggests the use of a framework by Lindblom (1993) for democratic governance of technology, called Intelligent Trial and Error (ITE). ITE poses 20 questions separated into five different categories. While assessing a project one can give the project a score from one to five for each question, sum the total score and divide it by 100. By doing this, one is left with a score between zero and one, representing how democratic the process of development or use is. Since the framework is of a generic nature, not specifically for AI or software development processes, there are a number of shortcomings. The questions are very high-level and while a user with experience and knowledge about the challenges of AI could reason their way from the questions to most of the relevant challenges, it may not be trivial for others who lack that expertise.

Clarke (2019) proposes 50 principles that participants in the various AI processes can use for guiding the process of designing, developing and using AI or evaluating to what extent they have ensured that the resulting AI is beneficial for society. The principles reflect most of the challenges I have identified, however, there are a number of shortcomings. First of all, as the principles are based on risk assessment and management, they fail to capture other aspects of democratisation not necessarily tied to risks, such as open access to data and technology produced by the project and how the

project and its artefacts should be governed. Further, the principles are not complete within the covered topics either. For example, they do not mention providing information regarding how reliable an AI system is or measures of negative bias in the predictions from the final model.

4.1 Decentralised control

While there seems to be a general agreement in the literature that the control of AI should be decentralised (Miikkulainen et al., 2019; Montes and Goertzel, 2019; Shah, 2018; Yigitcanlar and Cugurullo, 2020), the current rhetoric being used by China and other actors (Cave and ÓhÉigeartaigh, 2018) instead seem to reflect a situation where technological silos are competing for technological advantage or dominance. Further, there are already several examples of full-stack AI companies, such as Amazon, Facebook, Microsoft, Alphabet and Alibaba (Lauterbach, 2019). A full-stack AI company is a company that develops their own AI-optimised processors, host their own IT infrastructure, participate in fundamental AI research as well as develop and provide AI powered applications (Lauterbach, 2019). These companies tend to also cooperate closely with universities, compete to attract the most talented AI developers and buy up start-ups for their technology or talent (Lauterbach, 2019). Furthermore, full-stack AI companies tend to control a monopoly in some market, be it Google in search or Amazon in online retail, which helps explain how they fund the development of their technology stack (Lauterbach, 2019).

4.1.1 Challenges with centralised control

While there is not necessarily anything inherently bad with a company holding a significant grasp of AI technology, after all they may choose to act responsibly and use the technology in a way that is sustainable and beneficial for society as a whole. They may have a conflict of interest encouraging them to use the technology for their own economic benefit, while setting aside considerations regarding the consequences the actions may have for their users. Further, the fact that a company holds significant control over the technology does not line up with the principals of democratised AI, in particular the principle of decentralised control. The exception to this is the case where the company is adequately accountable to those who may use or be subject to the AI applications developed by the company, similarly to a representative democracy. This seems, however, not to be the case at present (see section 4.2.4) and I will discuss the challenges of accountability in section 4.2.

To prevent centralisation or monopolisation of the technology, one approach is to enforce legislation similar to the antitrust legislation used to prevent anti-competitive practices or monopolisation of a market (Ienca, 2019). An argument against splitting up companies that hold a monopoly within a segment is that the products would be less effective (Arogyaswamy, 2020; Shah, 2018). Shah (2018) instead proposes to tax monopolies according to how a competitive market would restrict their pricing. I will however argue that, while the taxing could be a good way of putting some of the money to work for society, it would not provide any check against the power

these companies wield in these positions and therefore would not mitigate their centralised control. My reasoning behind this is that taxing the companies would not necessarily affect how they make decisions regarding AI, perhaps except for incentivising them to exploit the technology further to make up for the money they are taxed.

4.1.2 Interoperability

A positive note here is that many of the AI tools and frameworks are available open source (Luce, 2019), and there are a number of AI platforms offered to the public by major stakeholders such as Google, Microsoft and Amazon (Masood and Hashmi, 2019). However, initially the development of open AI tools and frameworks were community-driven, but over time company owned or backed projects have taken over the majority of this space as companies have invested significant resources in developing their own solutions (Braiek et al., 2018). One potential explanation of this investment is an interest in centralising the control of the development of AI in their favour. By developing a popular framework for AI development and offering additional benefits such as large datasets and plug-and-play operations within their cloud infrastructure, they can capture significant control over the development as well as benefit from the large expenditures AI developers pay in order to train their models on the company's cloud infrastructure (Braiek et al., 2018). This effect can be further exacerbated by reducing interoperability, locking developers within their ecosystem by increasing the difficulty of migrating their code, models and data to other alternative services and frameworks. For this reason, the use of open standards for interfaces and other aspects of tools and resources can be an important tool in maintaining the flexibility to move between ecosystems (Belinchon et al., 2019; Miikkulainen et al., 2019).

This point also argued by Miikkulainen et al. (2019) and Belinchon et al. (2019), who argues that AI needs open standards and that interoperability is important to enable developers to build upon the previous successes of others and perhaps even more importantly, it discourages monopolisation of AI. On the contrary, Moreau et al. (2019) argue that components should only be standardised after they have been tested in a variety of contexts, and that the community prefers simple lightweight task-oriented components over interoperability, relying on simple file formats such as CSV and XML over standards proposed by Gate[7] (General Architecture for Text Engineering), Stanford CoreNLP[8] and Apache UIMA[9]. While these two arguments are not in direct opposition, both agree that there should be standards, they seem to disagree about the urgency of developing standards and encouraging interoperability. There are, however, several existing solutions for enabling compatibility between AI frameworks, such as the Open Neural Network Exchange[10] (ONNX) format for neural network models and Keras[11] as an abstraction layer to set up, train and run inference on NN

[7]Gate https://gate.ac.uk/
[8]Stanford CoreNLP https://stanfordnlp.github.io/CoreNLP/
[9]Apache UIMA https://uima.apache.org/
[10]Open Neural Network Exchange https://onnx.ai/
[11]Keras https://keras.io/

models (Masood and Hashmi, 2019).

4.1.3 Democratic governance

In addition control of resources it is also important to discuss the control of the development and use processes of AI. The literature did not specifically provide any approaches for democratic project management, but there were many authors writing about inclusive development, where people from the general public is included in the development and use processes which also promotes this notion of decentralised control (Ienca, 2019). These approaches will be discussed in section 4.5 about inclusiveness. However, one can draw inspiration from existing democratic models, such as participatory democracies, where the members participate and vote on each decision, and representative democracies, where the members elect someone to represent them and make decisions in their place.

Further, democratic governance also applies to state actors in order to ensure proper use and governance of AI on a national level, as well as applying appropriate pressure internationally to further the democratic AI practices. One can argue that state actors are held accountable through elections, which is true, assuming that the state is a democratic state based on rule of law. It is however not the case for states that suffer from corrupt or otherwise problematic electoral processes, do not have adequately informed citizens or are not democratic states to begin with.

4.1.4 Governing open resources

While the decentralised control of technology and data is a good principle, it does not come without its own challenges, such as anonymisation, trust, availability, integrity, ownership, load balancing, free-loaders and managing public access to petabyte-sized databases (Buckingham Shum et al., 2012). However, a centralised solution is not without challenges either, potential participants may not trust the central server and a central service becomes a single point of failure that could bring everything to a halt, potentially leak all the collected data if breached (Lyu et al., 2020) or be abused by the entity controlling it (Sudmann, 2020). In particular, there are situations where data is too sensitive to leave the place of origin, for example sensitive patient data from a hospital.

Buckingham Shum et al. (2012) argues that there are possible approaches for fair centralised control of a resource, suggesting the possibility of treating it like a common resource pool and to adopt an institutional management approach of it. To align this approach with the democratisation of AI, the institution controlling the resource would have to be democratically managed, thus creating a situation similar to a representative democracy. To maintain stability and fair use of the resource the authors refer to a set of conditions identified by Ostrom (1990):

1. Clearly defined boundaries to the resource and of institutional membership.

2. Congruence of provision and appropriation rules to the state of the local environment.

3. Collective choice arrangements are decided by those who are affected by them.

4. Monitoring and enforcement of the rules is performed by the appropriators or agencies appointed by them.

5. Graduated sanctions (i.e., more refined than 'one strike and you're out').

6. Access to fast, cheap conflict resolution mechanisms.

7. The right to self-organise is not subject to interference from external authorities in how the community chooses to organise itself.

8. That these self-organising institutions for self-governing the commons were part of a larger structure of nested enterprises.

One can also govern open resources through voting systems, where the access to vote and weight of one's vote can depend on one's contribution to the resource as a whole. It is of imperative importance that there are sufficient counter measures to keep malicious actors from artificially influencing the vote, either by creating a large number of accounts or boosting their account ratings (Montes and Goertzel, 2019).

When resources are openly available, as proposed in section 4.4 on openness, there is always a risk of people abusing them. For this reason, one needs methods of stopping such behaviour and protecting the resources such that they can be accessed by honest users. Fundamentally, there are two approaches to dissuade the misuse of an open resource, prevention and sanctioning (Buckingham Shum et al., 2012), where prevention attempts to keep the misuse from happening and sanctioning punishes misuse that has already happened.

Preventative measures can be distinguished into three separate categories: cryptographic techniques, obfuscation techniques and reputation techniques (Buckingham Shum et al., 2012).

Cryptographic techniques are used to make it prohibitively difficult or resource intensive for an attacker to gain access to a resource (Buckingham Shum et al., 2012). While encryption can be very effective, it also requires careful handling of encryption keys which becomes increasingly complex in situations where data and keys should be shared between a large number of actors in an open network instead of being centrally controlled. Hashing can be used to ensure the integrity of data in the network, but also requires careful handling such that an attacker may not substitute a malicious hash value.

Obfuscation techniques includes, for example, anonymisation of personal data or generation of synthetic data based on patterns in real data with the goal of making it impossible for someone to tie the data back to the origin, even if the attackers have access to other data and performs inference to deduce missing data points (Buckingham Shum et al., 2012). While it, by itself, does not keep an attacker from accessing the data, it protects those whose data is in the dataset if unauthorised access were to be gained. The downside of this approach is that valuable information can be lost in the process of obfuscation (Buckingham Shum et al., 2012).

Reputation techniques relies on the analysis of a user's previous behaviour in comparison to behavioural patterns previously observed (Buckingham Shum et al., 2012). While the approach may help detect malicious users, it does not guarantee to always produce correct conclusions and thus it may hinder some honest users or let some malicious users through the cracks (Buckingham Shum et al., 2012).

Sanctioning, on the other hand, is a retroactive approach in which a malicious actor is punished in some way in order to dissuade them from repeating their actions or deterring others from that type of behaviour (Buckingham Shum et al., 2012). Examples of this may be temporary or permanent bans from online platforms or restricted access to particular resources. A major issue with sanctions is that lacking identification of user accounts may allow users to simply create a new account, avoiding any sanctions put into effect (Buckingham Shum et al., 2012).

4.1.5 Research challenges for decentralised control of AI

- Improvement and development of new democratic governance methods to enable users to participate in the government of platforms and technology use, as well as the development of AI (Buckingham Shum et al., 2012).
- Buckingham Shum et al. highlights the need for establishing *ICT-enabled Governance* as a scientific domain with the focus on governance assisted with the use of 'formal methods, metrics and assessment models, decision support [as well as] simulation tools' (Buckingham Shum et al., 2012).

4.2 Accountability (and responsibility)

4.2.1 Ethical principles

The development of ethical principles for the development and use of AI is without question an important task. Ethical principles are agile, able to change at a rapid pace along with the development of AI technology and they can provide guidance to responsible organisations, corporations and other actors in the field of AI. AI ethics faces a few challenges, however. First of all, the principles are not enforceable (Lyu et al., 2020; Nemitz, 2018); an actor can simply choose to act irresponsibly, ignoring an ethical principle when it suits their interests, without any consequences beyond some criticism from interest groups or perhaps some public backlash, if they even reveal their practice to the general public. Secondly, ethical principles are inherently neither democratic nor global in nature (Belinchon et al., 2019). Thus, if the ethical principles agreed upon were created by people from a handful of countries and cultures, say for example the countries where AI is most prominently being developed, they would exclude important perspectives, principles and norms from most of the world (Belinchon et al., 2019). And there is some scepticism to whether it is actually feasible to come up with ethical principles that can be agreed upon globally or if it

is best to have several more localised ethical guidelines (Belinchon et al., 2019).

A number of AI companies have internal ethics boards (Lauterbach, 2019), which is a step in the right direction as they can provide better ethical oversight (Lauterbach, 2019; Shah, 2018). However, Google recently created an uproar by firing two heads of their AI ethics team in short order, potentially hinting at how there can be internal friction between the ethics board and the interests of the company. Lauterbach (2019) suggests that full-stack AI companies should be required to have transparent internal ethics committees to evaluate and ensure they are acting responsibly and that these committees should be as diverse as possible. Further, Floridi et al. (2018) argues that one should support the training of these boards, as well as making ethics boards mandatory for work on AI. On a similar note, the Norwegian Data Protection Authority (Datatilsynet) have created a regulatory sandbox, where companies can receive guidance to develop AI solutions that are ethical and responsible from a privacy point of view (Datatilsynet, 2020). This could certainly help, but it could not be considered as an alternative to enacting regulation of AI, as one still leaves the regulation entirely in the hands of the companies.

4.2.2 Regulation

In order to truly protect the rights and interests of the subjects of AI, one needs to establish enforceable laws and regulations that can be used to hold companies to account for their actions when they choose to act contrary to the interests of their stakeholders in the development and use of AI (Lyu et al., 2020; Shah, 2018). Having regulation to address relevant issues also enable authorities to perform audits to discover negative behaviour an actor may be been hiding from the public. Ethics can play an important role in the process of establishing these regulations, as regulatory bodies can use the ethical principles established in the AI community to help guide the regulation.

Baum (2017) highlights some challenges with using regulation to prevent certain features in AI design to ensure safe and beneficial AI. Namely that constraints may cause a backlash from AI developers or they may constrain the wrong features if they are not carefully considered. Further, the importance and impact of certain features in the design of AI may change over time such that the regulation regarding these features also will need to be kept up-to-date. Instead of restricting the features legal to use in AI designs, one could alternatively incentivise the use of beneficial features over harmful features (Baum, 2017; Djeffal, 2020). However, similarly to ethical principles, the developers could choose to ignore the incentives and go for a harmful design, despite any financial, public or professional backlash caused by the decision (Baum, 2017).

Instead of focusing on specific AI design features, perhaps it is more appropriate to address and constrain specific uses of AI on a higher level (Etzioni, 2018). The higher-level goal is after all not to constrain the features used in AI designs, but rather to ensure that the AI developed and used are beneficial for society and safe to use. Baum (2017) suggests to require the designer to choose the most beneficial design, but it does not

seem obvious that there is always a design that is objectively more beneficial than the alternatives, nor always a way to directly compare the benefit of different possible designs. Perhaps instead of requiring the designer to choose an imagined optimal design, the designer could defend their choice of design by demonstrating that all concerns and interests within reason were addressed and weighted appropriately in the process of choosing a specific design. This requirement of an appropriate process of reaching a design seems more feasible to demonstrate and evaluate than comparing a design to some imagined optimum.

There are also claims that laws would be too inflexible (Nemitz, 2018) or slow (Ghallab, 2019) to handle the future development of AI, but legislation has the ability to be quite general, technology independent, rather than specifically targeting AI or another technology specifically. Therefore it can instead focus on the primary principles and core issues that arise from the use of such technology. This is one of the core functions in a number of legal systems, for example in civil law systems, where general legislation is interpreted and made concrete through court cases.

Some authors even go as far as proposing the creation of an international governmental body for AI (Belinchon et al., 2019; Floridi et al., 2018), but others remain sceptical to the idea of creating governmental bodies specifically for AI (Belinchon et al., 2019; Shah, 2018) arguing instead that AI can be handled by existing bodies such as those responsible for data or consumer rights protection. Another alternative would be the creation of international multi-stakeholder fora where both policy makers, interest groups and other stakeholders could engage in a discourse to address concerns regarding AI (Belinchon et al., 2019). Challenges with open fora such as this includes deciding how to balance the influence of different groups regarding the discourse on specific issues and who should be allowed to vote on the outcome of different deliberations (Belinchon et al., 2019). One can, for example, imagine there would be fundamental disagreements about how to interpret equality, not to mention the challenges of aligning the views of different cultures or the influence of governments with competing agendas. However, that it may not be trivial to reach final decisions does not mean one cannot have useful discourses or outcomes. An example of a positive outcome from international deliberation, to reinforce this point, is the Universal Declaration of Human Rights.

Different countries have chosen different ways to approach the control and accountability of AI. In China, AI has become an important piece in the national strategy (Carter, 2020; Lauterbach, 2019), both as a way to surveil and control the population, but also as a step towards becoming a global technological leader (Schneider, 2020). The US, on the other hand, have chosen a free market approach where innovation appears to be prioritised over regulation of the technology (Schneider, 2020). Finally, the European Union have chosen to prioritise individual values and beneficial technology, thus creating initiatives such as the GDPR and the Ethics Guidelines for Trustworthy AI (Schneider, 2020). These different approaches seem to somewhat line up with the different views of equality, as previously discussed.

The next big question is then; which challenges of AI have to be regulated and which can be left to ethics? This is the point where literature

seems the most divided, and not without good reason. Regulating AI too much can lead to a slow down in technological development (Miikkulainen et al., 2019) and potentially drive research, development and investments to countries that are less strict, whereas regulating too little may lead to a situation where caution is thrown to the wind in a race to publish research first or getting first to market with a new technological development. Neither of these options are good for the future of our society, and as such, a balance needs to be struck.

Nemitz refers us to the principle of *essentiality* in legislation, meaning that any matter which concerns fundamental rights or the core interests of the state must be handled through democratic legitimised law (Nemitz, 2018). This means that issues regarding privacy, safety, autonomy, discrimination or the democratic processes cannot be left to self regulation by companies, all of which are issues that may be directly impacted through the use of AI (Ghallab, 2019).

The concept of fundamental rights as a foundation for the use, development and governance of AI seems to be supported by a number of authors (Donahoe and Metzger, 2019; Thinyane and Sassetti, 2020). First of all, the use of AI can benefit the realisation of human rights (Donahoe and Metzger, 2019), but the human rights can also inform the boundaries of how AI should be used. Thinyane and Sassetti (2020) proposes to use the International Bill of Human Rights as a foundation, highlighting that it contains additional information making it more suited for practical applications than, for example, the Universal Declaration of Human Rights. Thinyane and Sassetti importantly note that a system shouldn't just avoid violating rights, but that the developers and users should respect and consider the rights throughout the development and use processes. Further, they suggest following an approach by Raso et al. (2018) to evaluate the human rights impact of an AI solution, focusing on the quality of the data, the system design and the inclusiveness of the process, and finally the interactions between the system and the context into which it is deployed (Thinyane and Sassetti, 2020).

Nemitz (2018) argues that there is a need for regulation to enable transparency on three levels. First of all, they propose that an impact assessment is performed for policy making and legislation before deployment of high risk technologies, such that adequate legislation can be developed in order to protect essential interests. Secondly, they argue that developers and users of AI should perform impact assessment in regards to whether the use can impact democracy, rule of law or fundamental rights. And that the assessment should be publicly available if the AI will be used in the public, and that public authorities should perform their own independent assessment in high risk cases. Finally, they argue that AI users should be legally compelled to provide an explanation of how the AI works and how it impacts individuals. While some of these already apply under GDPR when it comes to algorithms processing personal data, the authors propose extending the regulations to a wider set of aspects including democracy, rule of law and fundamental rights, as mentioned earlier.

In a very recent development, the European Commission published a proposal for a regulatory framework on AI (European Commission, 2021). The framework divides AI applications into four groups based on risk, rang-

ing from minimal to unacceptable risk. Applications that pose a clear threat to fundamental rights of citizens are deemed unacceptable and unlawful. Further, applications that deal with safety, asylum and migration, remote biometric identification, as well as access to education, employment and essential services, fall within the category of high risk. These applications will face a set of strict requirements, such as risk assessment and mitigation, detailed documentation of compliance, high quality datasets as input and high levels of robustness. The limited risk category includes applications that requires some level of transparency, such as how users must be informed that they are dealing with an AI agent instead of a human being. And finally, the minimal risk category covers all applications that have no legal obligation.

4.2.3 Unemployment

Unemployment as a result of automation is a big challenge in the face of AI (Belinchon et al., 2019; Ghallab, 2019; Lauterbach, 2019; Wirtz et al., 2019). The challenge is not a new one, as some of the most prominent examples of it dates back to the industrial revolution. However, AI has the potential to cause massive layoffs due to the automation of professions where such processes were not previously feasible (Lauterbach, 2019). The question here is where does the responsibility for this issue lie? Is it up to the people laid off to re-educate themselves and find new work, should the state take responsibility by providing support for those affected (Belinchon et al., 2019; Floridi et al., 2018), or should the companies who automated the positions be held accountable? The answer to this question seems to be tied to how one interprets equality and labour rights, as previously discussed.

4.2.4 Approaches for accountability

There are essentially three methods of holding actors accountable regarding the democratisation of AI, through capitalistic mechanisms, regulatory mechanisms and political mechanisms. The capitalistic mechanism can be used to hold capitalistic actors accountable through alternatives in a market, where users can choose to move to an alternate service if they disagree with the actions of their current service. It relies on companies regulating themselves as to not lose users to competing services. There are some authors who argues for relying more or less entirely on such an approach (O'Sullivan and Thierer, 2018), where the regulation of the technology is light or nonexistent and instead one should trust the companies to do what is right. This mechanism runs into problems, however, when there are no legitimately competitive alternatives for users to switch to or when users do not act in their own best self interest. The issue of competitive alternatives proves to be a particular problem when it comes to actors involved in surveillance capitalism, where the accumulation of data about the users enables the actors to make large sums of money from the advertisement industry. As of April 2021, the top 5 most valuable companies in the S&P500 (Apple, Microsoft, Amazon, Facebook and Alphabet, parent company of Google) were companies with advertising networks, making out over 20%

of the total value in the S&P500 index. This speaks to the challenge other companies face when competing with the added revenue from selling direct advertising. Even when assuming that there is a competitor who manages to establish a similarly featured service without resorting to surveillance, there is also a notion of inertia keeping users in their current service. They are already familiar with their current service and probably have quite a lot of data stored there already, not to mention social media where users would essentially have to convince their contacts, jobs, organisations, universities and so forth, to swap service in order not to lose touch with their friends and to stay up-to-date with their events, plans and relevant announcements. Further, all of this additionally relies on well-informed users that act rationally in their own best interest, that the service is transparent and honest about their actions and that each user has enough knowledge on the relevant topics to actually understand the severity of the service's actions. In other words, there are a lot of obstacles in keeping companies accountable through capitalistic mechanisms and the suggestion that this is an effective mode of regulation when it comes to AI seems naive. Other authors have also highlighted self-regulation as an approach with limited effect, whether it is due to a conflict of interest, focusing on metrics such as shareholder value instead of impact on users and society, not realising the potential consequences of their decisions or not taking the consequences serious enough (Arogyaswamy, 2020).

The regulatory mechanism is to ensure accountability through regulation enforced by a government. As argued previously, regulation might be the most important tool to ensure AI is not abused within a country, due to how there seems to be a lack of responsiveness in the capitalistic mechanism. For example, data protection law can be used to regulate how a user's data can be utilised for AI development (Djeffal, 2020). But beyond setting restrictions on how AI can be used, law can also encourage development of certain forms of AI (Djeffal, 2020). Failure to properly regulating how AI is applied can lead to a normalisation of breaching fundamental societal values and rights, such as how privacy and autonomy seems to be washed away by surveillance capitalistic actors (Zuboff, 2019). It can pose a particular risk for social groups that are already disadvantaged, such as historically-marginalised groups (Yigitcanlar and Cugurullo, 2020). A number of authors dispute that AI should be regulated until it is absolutely necessary, arguing that it would stifle the creativity and slow down the development of the technology (Miikkulainen et al., 2019). Others argue for a compromise, where industry and government co-regulate the technology as governments are less suspicious of the companies and the companies still can, to some extent, control the regulation of the technology (Carter, 2020). Companies already seem to hold significant power over government regulation, such as through lobbies and dark money schemes (Hertel-Fernandez et al., 2018). And this control seems, in many cases, to come at the expense of the general public, for example by protecting tobacco advertisement (Neuman et al., 2002) and fossil fuel subsidies (Victor, 2009). In 2018 Google was the company that spent the most money on lobbying in the US, spending more than $21 million (D'Souza, 2019) to fight off privacy- and other regulation that would hamper their collection of user data or otherwise interfere in their efforts.

Thirdly, the political mechanism is used to enforce accountability internationally. This can include using sanctions such as tariffs or political pressure to dissuade countries from breaching international agreements. OECD (Organisation for Economic Co-operation and Development) created the first international accord for AI (Carter, 2020). While the accord was accepted by many countries, China decided not to follow the principles proposed in the accord, highlighting instead how the country views AI as an technological race and they are intent on being the leader in the field (Carter, 2020; Lauterbach, 2019). Despite their competitive rhetoric, China mentions in their own guidelines that they will work to improve international cooperation when it comes to laws and regulation of AI (Carter, 2020).

Research objectives tied to AI and accountability includes:

- Develop regulatory models that are effective for governing emerging AI activities and capabilities (Smith and Neupane, 2018), these new models must also be able to keep up with the rate of innovation in the field.
- Research whether more regulation of AI will actually impact the rate of innovation in the field and alternatively how one can regulate to protect values and interest with the least possible impact on innovation.
- Develop systems for determining accountability when an AI misbehaves and makes erroneous decisions (Smith and Neupane, 2018) that are discriminatory, causes financial loss or even harm to a person.
- Investigate how AI can support or diminish human rights (Smith and Neupane, 2018), either by enabling more people to achieve a right or violate rights such as privacy or autonomy.
- Investigate AI's impact on the job market and how to achieve a future where AI development and use does not come at a great cost for employees subjected to job automation (Smith and Neupane, 2018).

4.3 Transparency

4.3.1 Algorithmic transparency

The literature seems to agree that there is a need for algorithmic transparency, insight into the logic behind algorithms, in such a way that it can be understood by those subjected to the algorithm (Carvalho et al., 2019; Djeffal, 2020; Floridi et al., 2018; Kobayashi et al., 2019; Moreau et al., 2019; Timan and Grommé, 2020). The term algorithmic transparency incorporates the concepts *interpretability* and *explainability*, where interpretability refers to the extent to which one can interpret the inner workings of algorithm itself and explainability is to what extent one can understand and explain how an algorithm reached some conclusion (Baird and Schuller, 2020). The two terms are, however, often used interchangeably in the literature and there seems to be a variety of definitions for the terms in the literature (Carvalho et al., 2019; Hohman et al., 2019). Carvalho et al.

(2019) states that algorithm transparency is more about how the algorithm learns from data and what relationships it can learn, rather than the model the algorithm learns. Thus it seems there is either a disagreement in the literature regarding what the term algorithmic transparency should mean, or algorithm transparency and algorithmic transparency are two distinct terms that should not be confused. For the purpose of this thesis, however, I shall use algorithmic transparency as a broader term which includes the learned model, as defined above.

Gaining the insight of algorithmic transparency is of major importance for critical systems, especially relating to health and safety, as the failure to explain why wrong decisions were made may result in a lapse in trust and may cause resistance to further developments (Gao et al., 2018; Yigitcanlar and Cugurullo, 2020). Further more, it could facilitate or improve the general public's right of explanation for automated decisions made about them, as required by the GDPR. Algorithmic transparency is also useful to address issues regarding negative bias (Hohman et al., 2019), as getting insight into how the decision was made could help establish whether skin tone, gender or other sensitive attributes had a significant impact on the outcome. Further, algorithmic transparency would be useful in enabling public discussions about the outcomes of an algorithm. It could enable authorities to demand explanations for certain behaviours of proprietary systems without the owners of the systems needing to reveal the inner workings of the algorithm and thus risk exposing their intellectual property (Carvalho et al., 2019). Finally, algorithmic transparency can be used in education to help illustrate how the algorithms work internally (Hohman et al., 2019). However, while useful, it seems not to be strictly necessary for legal accountability, as the legal system already have methods of handling black-box processes (Wischmeyer, 2020).

The issue of algorithmic transparency is particularly relevant for neural networks, due to the nature of how they function. There does however already exist a number of approaches to reveal information about why or how a neural network reached a conclusion. In Local Interpretable Model-Agnostic Explanations (LIME) one changes the input slightly and see how it impacts the output of the neural network (Shah, 2018; Sudmann, 2018), while it does not reveal any natural language arguments for an output it can reveal to what extent each part of the input contributed to the output. This delta can be analysed by a domain expert who can evaluate the relevancy of the features highlighted and point out cases where the highlighted features can be tied to negative bias in the algorithm. But while this is easy for simple queries, it may become non-trivial for complex inputs. Such as when none of the input features are direct representations of a minority but instead the input contains a collection of proxy features that may be combined to represent certain minorities such as address, income, name, profession, etc. A Pointing and Justification (PJ-X) model can be used to provide predefined answers to specific questions regarding the input, such as why a person in an image is not flying (Sudmann, 2018). D. Wang et al. (2020) proposes the use of visualisation tools to illustrate the technical details of the algorithms resulting from AutoAI. AutoAIViz by Weidele et al. attempts to address this need, visualising the process of selecting model components and parameters. Carvalho et al. in their work

Machine Learning Interpretability: A Survey on Methods and Metrics (2019) provides a great overview of different approaches and perspectives for algorithmic transparency that exists in current literature. Further, Hohman et al. (2019) offers an overview of use cases and existing solutions for visualising deep learning.

While algorithmic transparency is an important avenue of research, Wischmeyer (2020) also highlights that regulatory agencies do not have the resources necessary to process all the data and information already provided regarding the AI systems in use. Thus, there is also a need to provide these agencies with additional resources or simplifying the data needed for a review such that less resources are required.

4.3.2 General transparency

Instead of full algorithmic transparency, the focus could alternatively be on gaining insight into the information surrounding a decision made by an AI system such that the decision could be changed or that the relevant stakeholders can decide accountability in that particular context (Wischmeyer, 2020). This seems to be a useful alternative, at least while more sophisticated approaches for algorithmic transparency is being researched.

Another challenge is ensuring that users and other stakeholders have adequate information such that they can make informed decisions regarding and trust the use of an AI system for a particular use case (Wischmeyer, 2020). But one has to find a balance between the need for the general public to have access to information regarding the systems that make decisions about them and the legitimate interest of the companies developing these systems to protect their technology and intellectual property (Wischmeyer, 2020). Exactly what information is disclosed may depend the context of the system (Wischmeyer, 2020), such as whether it is run by the public or private sector, what information does the controller have a legitimate interest in keeping secret, is the system decisions critical in nature, etc.

To enable transparency for users and other stakeholders, one could first of all disclose whether or not AI is used in a system (Wischmeyer, 2020). The GDPR gives the subjects of an automated decision the right of explanation for the outcome of that decision. This right could be extended by giving the general public the right to a set of information regarding systems and services, such as whether AI is used or not, even if they are not subject to that system (Wischmeyer, 2020). One can imagine this information provided on similar grounds as the privacy policy of a service.

Machine Learning algorithms are inherently probabilistic and while the accuracy, precision and F1 scores are prominently displayed for benchmark datasets when new algorithms are published, it is also important to surface it when the algorithms are put into practice. To make the information more digestible for non-experts the metrics could take a different form such as a confidence score or other metric to communicate the reliability of the algorithms. These metrics and scores could be calculated using standard processes which may depend on the use case and type of algorithm. If services provide a confidence score for their algorithms, users can decide whether the confidence is appropriate for their use case, or if their use case requires an even more robust algorithm (Baird and Schuller, 2020; Floridi

et al., 2018). A confidence score could, for example, be calculated with some formula to include measures of bias for relevant minority groups (Baird and Schuller, 2020), precision scores, accuracy scores and other relevant measurements that may inform the robustness of the algorithm.

Rather than providing these scores and details in a large document, as has been the tradition with privacy policies, instead Belinchon et al. (2019) argues for the use of labels on products and services that would inform users about the collection of data and use of machine learning. These labels could also contain the confidence score of the model and information about to what extent it was checked for negative biases (Shah, 2018) as well as other metrics that may be relevant. There exists already a number of initiatives to produce labels such as these, however primarily focused on privacy rather than AI, including ToS;DR[12], Privacy Labelling (Johansen et al., 2021) and Apple Privacy Labels[13].

Finally, governmental agencies could require access to a deeper set of information regarding an AI system than the general public on the premise that the information will remain secret. Thus AI systems need to be auditable, in particular the data used for training needs to be reviewed and be auditable to ensure the AI has a sound foundation (Baird and Schuller, 2020; Timan and Grommé, 2020). Further, the processes used in the design and development of AI should be auditable to ensure appropriate steps are taken to mitigate negative bias and potential unfair treatment of minority groups (Floridi et al., 2018), as argued in section 4.2.2.

Research objectives for AI and transparency includes:

- Develop methods of uncovering bias in AI applications before they are used in public spaces (Smith and Neupane, 2018)
- Investigate what scores and metrics are relevant to disclose to the public when publicly deploying AI services
- Explore the feasibility of a confidence score and potential processes for how to derive it
- Develop labels for AI scores or extend existing privacy label systems to include AI information
- Investigate where the boundary lies between disclosing information for the general public and protecting the legitimate interests of the algorithm developers

4.4 Openness

4.4.1 Dual-use

While there is an agreement that openness is core to the democratisation of AI (Allen et al., 2019; Belinchon et al., 2019; Holford, 2019; Ienca, 2019; Miikkulainen et al., 2019; Shah, 2018), there is a dispute regarding to what extent this should happen (Ienca, 2019; Miikkulainen et al., 2019; Sudmann, 2020). On one hand, keeping the technology open and available

[12]ToS;DR https://tosdr.org/

[13]Apple Privacy Labels https://www.apple.com/privacy/labels/

enables a wider community to apply it for use cases not considered by the original researchers or developers, as well as for other researchers to reuse existing solutions rather than reinventing the same technology (Moreau et al., 2019). On the other hand, since AI technology in many cases can serve more than on purpose, it can enable people to misuse the technology for nefarious purposes, be that illegal operations or in military applications (Cave and ÓhÉigeartaigh, 2018; Sudmann, 2020). This dual-use aspects highlights the need for there to be ethical assessments in the planning, development and use AI systems, as well as regulation to make it possible to hold those who misuse AI accountable for their actions (Ienca, 2019). Further, this issue raises the question of where the boundary between beneficial and dangerous AI.

To begin tackling this dual-use aspect of AI Lauterbach (2019) suggests that scientific institutions could act as trustees of datasets and algorithms to ensure they are used responsibly and for good purposes. Further, the author proposes that potential negative societal consequences of research should be disclosed in the published work. Increasing the awareness and consideration regarding the potential for misuse in the development of AI sounds like a good first step towards more beneficial technology. Indeed, other authors mention that AI researchers could benefit from additional education in ethics and social sciences to increase this awareness and forethought. However, the suggestion that scientific institutions should safeguard technology does raise some potential issues. For example, how does one decide what technology needs to be controlled? Who should hold these institutions accountable to ensure they balance the openness and safety concerns properly? How does one decide whether another party can be trusted with technology that could potentially be misused?

4.4.2 Race for AI

A race dynamic for the development of AI technology has to be avoided as it may cause participants to cut corners in regards to safety, ethics or regulation, leading to an increased risk of solutions that can be to the detriment, rather than benefit, of our future (Cave and ÓhÉigeartaigh, 2018). Lauterbach (2019) argues that there cannot be any notion of a race in AI development as there is no finish line. I will instead argue that there can be a race between two or more parties, even if the finish line is not formally established, as long as the parties are set on being ahead of the others on the same path. In the case of China's competitive AI rhetoric, the path is the development of AI technology. And while there may not be a formally established finish line to decide winners and losers, the potential race is toward any AI developments that would benefit them in some way.

However, some authors actually highlight a potential goal for an AI race: the super intelligence (Baum, 2017). An Artificial General Intelligence (AGI), is a hypothetical AI algorithm that is capable of learning a broad spectre of different behaviours (Muehlhauser and Salamon, 2012), similarly to humans, in contrast to the narrow AI of today which are highly specialised in one particular task. An AI super intelligence extends this by also surpassing human capabilities in the related tasks, which could hypothetically lead to something called the intelligence explosion (Muehlhauser

and Salamon, 2012). A situation where an AI is capable of developing new algorithms better than itself and thus recursively increasing the technological capabilities of those controlling the AI (Muehlhauser and Salamon, 2012). The concern is that the actor who controls this AI would rapidly outpace the research efforts and capabilities of any other powers keeping them in check. It is important to note here that it is in fact irrelevant whether artificial super intelligence is, or will ever be, possible. Having super intelligence as the goal of an AI race only depends on the participants believing that they need to reach this technology before the others.

Further, a race dynamic does not only apply to nation states, as there could also be negative competition between research groups or companies in getting their results first published or first to market. This further emphasises the need for regulation in order to control what corners participants can cut in the race to their results, as discussed in section 4.2.2.

Perhaps openness and cooperation regarding results and technology among research communities, and states, as well as a focus on safe and responsible development of AI for the global good, could be practised to maintain trust between the parties (Cave and ÓhÉigeartaigh, 2018) and promote beneficial AI instead of harmful competitive practices (Baum, 2017). There are, however, authors disagreeing with this, instead arguing that openness in AI may actually contribute to the race dynamic, in particular in proximity to the goal of a super intelligence (Bostrom, 2017). Their reasoning for this, is that in a close competition there would be less room to slow down and ensure one makes safe decisions (Bostrom, 2017). Actors might also be less willing to choose options that reduces the effectiveness of the AI in trade for more safe operation (Bostrom, 2017). On the other hand, a closed environments with technological silos competing for the advantage does not necessarily alleviate these problems. The silos may still choose to charge ahead with a disregard for safety, just in case the other silos are right behind them. One could also end up in a situation where a single company or nation is in control of technology which is vastly superior to the rest of the world. In an open environment, on the other hand, more actors will be at the same level of AI development and thus will not be able to overwhelm the others.

Cave and ÓhÉigeartaigh (2018) identifies three sets of risks when it comes to a race for advantage in the development of AI technology. First, the authors highlight the risk of a competitive rhetoric. Secondly, the risk of there being an actual race for AI. Thirdly, the risk of someone winning an actual race for AI. If the actors engage in competitive rhetoric, they may not be able to develop broadly beneficial AI as their rhetoric may eliminate dialogue and collaboration towards the necessary agreements, goals and compromises (Cave and ÓhÉigeartaigh, 2018). Further, the use of competitive rhetoric may naturally spark an actual race of AI (Cave and ÓhÉigeartaigh, 2018).

A competition in AI development may lead to the participating actors cutting corners regarding safety, ethics or other important considerations in favour of getting their results as quickly as possible (Cave and ÓhÉigeartaigh, 2018). It may also lead to conflict between the actors, such as espionage, sabotage or in the worst case an open conflict (Cave and ÓhÉigeartaigh, 2018).

There seems to be some major roadblocks on the path to achieving open international cooperation on AI development, however, such as political animosity between states and ethical disagreements such as how some states might use the technology for surveillance of their citizens or autonomous lethal weapons. One may have to look to historical examples of successful international cooperation for inspiration on how to approach such an agreement rather than resorting to secrecy. For example the International Space Station, World Trade Organisation, the UN climate panel and other multinational collaborations about establishing common standards. However, excluding certain states from the cooperation might be taken as an excuse to return to a technological race dynamic.

4.4.3 Access to data

Part of openness of AI is ensuring open access to data (Musikanski et al., 2020) and other relevant resources for the general public. Academic data resources and datasets are often shared as they are expensive to produce, but software prototypes are are typically not published even if they are sometimes reused to achieve new goals (Moreau et al., 2019). Beyond academic datasets, there are other cases where there is not enough data available, often due to privacy concerns, such as in medicine (Kobayashi et al., 2019; Luce, 2019). A cost-saving alternative for less sensitive scenarios is to engage non-experts in the creation of data (Moreau et al., 2019), for example by involving a community in playing a video game as was done for the MineRL project (Guss et al., 2019), however there is a need to develop more methods of enabling this (Moreau et al., 2019).

From the literature I have identified a number of different ways of managing data in an open fashion. The control of user data is of particular importance in this time of surveillance capitalism, notably that a user should have awareness, accessibility, consent and reusability of the data (Buckingham Shum et al., 2012; Musikanski et al., 2020; Timan and Grommé, 2020). One way to tackle this is by creating a decentralised marketplace where the data of each user is aggregated and they can control what purposes it can be used for Belinchon et al. One could imagine that a user accepts that a service may add information about the user to this database and the service in turn would get a percentage of the user's earnings from the provided data. In this variant the user maintains control of the data and the service can still make a profit from gathering data about the user, granted that the user actually sells access to their data. This still does not solve all problems, as the control of exact knowledge about what data is mined about users and how it is processed is an important market advantage for companies based on surveillance capitalism, thus one can imagine they will be very reluctant to give up this control and exclusive knowledge. Further, it is still surveillance capitalism, and relies on companies gathering copious amounts of data about their users. Therefore there should be alternatives for the users such as plans where you pay a subscription rather allow the company to gather the data. This concept seems to line up with existing regulation, such as the GDPR, granting users control over their personal data, but extended to cover all data associated with the user even if the user's account or ID cannot be associated with a natural person.

Shah (2018) proposes a somewhat different approach, where the company would steward the data for a limited time, after which the control of the data would be released to the public domain. I see three issues with this idea, however. First of all, it is only concerned with that the data will be openly accessible, not considering how a user should be able to control their own data. Secondly, it only applies to non-personal and anonymised data, personal data would have to be handled the way it is today as most users might be opposed to their data being shared openly. Thirdly, some data is time sensitive, such as current location or search queries, after the stewardship period ends the data may have reduced value or be useless.

There are also approaches that would enable the publishing of datasets that contain personal data. Braiek et al. (2018) proposes the use of generative models to train on massive existing data sets and then using the models to generate large synthetic data sets based on the original data. These generative models would also be easier to transfer between interested parties as they take up a lot less space, and thus also consumes less resources of the infrastructure where the dataset is hosted. There is however an upfront cost of resources to train the model on the original data. Typically, a generative model refers to the generative side of an Adversarial Generative Neural Network (AGNN), where a generative network generates, for example, an image and a discriminatory network attempts to discern whether the image was generated or is part of the original dataset. The generative network is rewarded when the discriminatory network believes the image it generated was from the dataset, and the discriminatory network is rewarded when it correctly concludes the origin of the image. Over time, the generative neural network will learn how to generate images that are practically indistinguishable from the original dataset. While generative models can address a number of challenges, the users of the resulting models will need to keep in mind that the generative model may reflect, or in the worst case exacerbate, biases existing in the original dataset (Jain et al., 2018).

Montes and Goertzel (2019) argues for the creation of large decentralised datasets where privacy information regarding each point of data can be decided by the contributor. This idea of creating a large decentralised dataset is also suggested by Belinchon et al. (2019) and Baird and Schuller (2020), where Baird and Schuller also highlights possibilities for using homomorphic encryption, data masking and federated learning as ways of preserving privacy for the data owners.

In some cases, the data is too sensitive to be moved around, but there is not enough data stored locally to train robust AIs. For this purpose Lyu et al. (2020) proposes a learning framework, called Fair and Differentially Private Decentralised Deep Learning (FDPDDL), that uses blockchain to enforce a decentralised reputation system on the users of the framework and enables local training of AI. The reputation system keeps track of the contribution of each participant and restricts them from extracting considerably more value from the framework than they contribute. Further, the framework employs Differentially Private Generative Adversarial Networks (DPGAN) and Differentially Private Stochastic Gradient Descent (DPSGD) to enable the network to move a model around, training it on data locally, rather than moving data around and potentially risking exposure.

4.4.4 Data exploration

Even though a dataset may be open to the public, it may not be trivial for others to put it to use as the documentation of the data is often not sufficient to fully understand the data and its context (Choi and Tausczik, 2017). To this end, a greater focus on properly documenting public datasets and collaboration with domain experts in interpreting and analysing the data can help clarify the details (Choi and Tausczik, 2017).

To further understand the content of a dataset one can utilise data visualisation tools to explore what data is available and the biases existing in the data. Data visualisation tools with a high usability can enable domain experts to participate in the analysis of the data (Gao et al., 2018) and thus visualisation tools can play an important role as an enabler of inclusiveness. Efficiency is a major challenge for these tools (Gao et al., 2018) due to the potential size of the datasets in use for AI. However, Kraska (2018) found that query approximation, progressive results and an analytical workflow that is easy to change was more important to interactivity than having a fast execution engine.

One potential challenge when enabling non-experts to work with statistics and analytical tools without any previous training, is that they may make mistakes such as drawing conclusions from insignificant changes in results. To help users avoid such mistakes, Kraska included a component called QUDE (Quantify the Uncertainty in Data Exploration) in their Northstar data science system (Kraska, 2018). The inclusion of error avoidance solutions may play an important part in enabling a wider community to analyse data and develop AI, as issues such as statistical insignificant results could lead users to draw false conclusions.

4.4.5 Auto ML and hardware access

Hyper parameter tuning is a particular challenge for machine learning (Gao et al., 2018). Finding the optimal parameters for learning rate, training algorithm, number of iterations and batch size, not to mention what algorithm and architecture to use, can be both tricky and time consuming, as there is typically a very big number of parameters to tune and changing one parameter can change what the optimum is for other parameters. If one is attempting to solve a known problem, such as image classification, there are often several well-performing models to choose from. The challenge in this case is that they each have their own trade-offs, typically resource intensiveness versus accuracy (Gao et al., 2018), thus deciding which one to use requires experience that most non-experts might not have. The problem becomes even harder when there are no standard models to turn to, such as in the case of novel AI problems and applications. Further, to determine how a combination of parameters performs requires training a model on the dataset, which can take anywhere from minutes to years, depending on the problem and hardware. Reinforcement Learning (RL) is particularly notorious for requiring vast amounts of training time, where projects such as AlphaGoZero and OpenAI Five used more than 4.9 million games of Go and 11,000 years of Dota 2 game time, respectively (Guss et al., 2019).

This challenge highlights two particular needs. First and foremost, one

needs access to hardware capable of the computation necessary for training the model and tuning the hyper parameters within a reasonable span of time. For this, hardware in the cloud can be a solution (Luce, 2019). However, cloud compute and cloud storage is rather expensive for people who are not using it professionally (Luce, 2019). An alternative, would be taking inspiration from the BOINC [14], Folding@Home [15] and Leela Chess Zero [16] projects, enabling peers to donate compute time on their devices to massively distribute large computing tasks. Further, one can develop more efficient algorithms that requires less computing to achieve similar results. K. Wang et al. (2017) proposes an algorithm they claim is easier to use, as one does not need to specify the learning parameters beforehand, as well as requiring less compute time in comparison to gradient descent.

Secondly, there is a need for software to handle this hyper parameter tuning automatically (Gao et al., 2018), keeping track of previous results, models and parameters as well as using automatic analysis to decide what parameter combinations to try next. These types of solutions are typically referred to as Auto ML and they can have a varying degree of complexity, from automatically comparing a set of parameters, to automatically configuring data pipelines, selecting algorithms and choosing an architecture and hyper parameters for the problem.

When it comes to the selection of algorithms and architectures Auto ML (or Auto AI) can be used to help select a reasonably effective model for simple problems, but requires further research to truly become usable for a greater variety of real world problems. It could serve as a tool to enable a wider public to develop AI for themselves, without deep knowledge in AI, and many papers highlights this as a good place to start (Allen et al., 2019; Bagrow, 2020; Binnig et al., 2018; Masood and Hashmi, 2019). More complicated problems, on the other hand, might require collaboration with AI experts to come up with an algorithm and architecture suitable for the problem.

Bagrow (2020) performed an experiment to show that non-experts are able to come up with prediction tasks, of their own interests, that are trainable by Auto ML algorithms. This finding is a positive note indicating that non-experts can indeed participate directly in the development of AI with the current Auto ML technology, not only by assisting with surrounding tasks such as data collection or labelling.

However, D. Wang et al. (2020) recognises two challenges in the effort of enabling non-AI-experts to make use of AutoAI, *Business Objectives* and *Transparency*. Firstly, they define Business objective challenge as being the tension that is created by that most business objectives desired by the non-technical users, for example compliance to regulations, often do not align with the technical objectives, such as accuracy (D. Wang et al., 2020). To address the business objective challenge D. Wang et al. (2020) proposes the use of constraints as an approach to not only satisfy the technical objectives but also the business objectives in the AutoAI process. Secondly, they define the transparency challenge as the challenge for non-expert users to truly

[14]BOINC https://boinc.berkeley.edu/
[15]Folding@Home https://foldingathome.org/
[16]Leela Chess Zero https://lczero.org/

understand how the AI works and why certain outcomes happen. This challenge lines up with the reasons for why algorithmic transparency is such an important avenue of research. Algorithmic transparency is discussed in section 4.3.1.

There already exists quite a few Auto ML solutions, both services such as Azure ML[17], Amazon SageMaker[18], Google Cloud AutoML[19] and Weights & Biases[20], and libraries, for example Neural Network Intelligence[21], Scikit optimize[22] and Keras Tuner[23]. The literature also includes a variety of solutions to address this challenge:

- Kojima et al. (2020) developed Kyoto-university Graph Convolutional Network framework (kGCN) an Auto ML system for working with chemical structures.
- Olson et al. (2018) developed a system named PennAI, an Auto ML system for biomedical data which helps guide the user through the process of deciding which models to try on the data and what parameters to change.
- Liang et al. (2019) developed LEAF (Learning Evolutionary AI Framework). It uses evolutionary algorithm to not only select model parameters, but also network architectures.
- Shang et al. (2019) presents an Auto ML system called Alpine Meadow, focused on interactivity, small data and traditional ML pipelines (as opposed to Neural Network architectures).
- Binnig et al. (2018) has proposed a system called Quality-aware Interactive Curation of Models (QuIC-M) that automatically constructs AI data processing pipelines and curates a model based on a user-specified problem and constraints.
- Völcker et al. (2020) have developed a system named DeepNotebooks, which takes a tabular dataset and produces a model and a python notebook for interacting with the model. The user can then further refine the parameters used and query the model and data using python.

To help evaluate all these different solutions, Xanthopoulos et al. (2020) created a set of criteria to evaluate Auto ML systems and services from a user's perspective, rather than focusing on the performance numbers of different solutions. The 32 proposed criteria are spread across six categories, including estimates, scope, productivity, interpretability, customisability and connectivity. Some of these criteria line up with challenges discussed in this chapter, highlighting features of the services that support the democratisation process. For example, interpretability aligns with data visualisation and algorithmic transparency, and connectivity relates to interoperability and AI access.

[17]Azure ML https://azure.microsoft.com/en-us/services/machine-learning/automatedml/
[18]Amazon SageMaker https://aws.amazon.com/sagemaker/
[19]Google Cloud AutoML https://cloud.google.com/automl/
[20]Weights & Biases https://wandb.ai/
[21]Neural Network Intelligence https://github.com/Microsoft/nni
[22]Sci-kit optimize https://github.com/scikit-optimize/scikit-optimize
[23]Keras Tuner https://github.com/keras-team/keras-tuner

4.4.6 Data preprocessing

A lot of the time of an AI developer goes into the preprocessing of data, or data wrangling, which is the development of pipelines to take the raw input data, massaging it into a format that is suitable for AI training and extracting the features in the data that the designer consider relevant for the prediction task. Patel (2020) suggests that AI features can be democratised through feature stores, databases where one can get access to already extracted features and information about the features.

I would like to expand upon this idea with the addition of data wrangling pipelines, where the extracted features are not stored directly but instead a pipeline that one can feed the data through to extract the features in the format one needs. These pipelines could be made modular, where one inserts modules one after another to add layers of processing. This modularity would in particular be useful for people with less programming expertise, as they could design their own pipelines though graphical user interfaces, relying on existing pipeline modules developed by others in the community. With such a system others could take existing pipelines and alter them to their own needs, for example someone might want pictures in black and white rather than colours or perhaps scaled down to a different size, depending on their use case. This approach has other advantages, namely that one does not need to store data multiple times, first in the raw form and then a copy in each other format that is useful to the community. The downside is of course that the final dataset needs to be computed from the raw data for each particular case, but each user would only need to do this once to create their own dataset. If more data is added to the raw data store one only needs to process the new data, requiring minimal additional work.

4.4.7 AI access

After the AI has been developed, it needs to be tested and validated for its purpose, deployed and the deployment needs to be monitored to ensure there is enough capacity to handle incoming traffic and that crashes are appropriately handled (Gao et al., 2018). Model validation is about ensuring that the model performs as expected on unseen data, not just on the training dataset. For example, while a trained model may perform well on the dataset as a whole, there may be subsets of the data where the model is performing significantly worse. An example of this may be an AI face detection system that gets confused by different skin colours. Chung et al. (2019) have developed a system for discovering such sub-performing subsets of the dataset with the goal of highlighting potential problem features that the developers should consider taking a closer look at.

After a model has been successfully validated, it needs to be set up for deployment. This could, for example, be done through a web API or as a software package that can be integrated into an application. To make this process more approachable for a wider community, these functions could be implemented as a service that could be set up and rolled out on demand (Gao et al., 2018). There are already a number of such services offered by companies such as Google, Microsoft and Amazon. The literature provided

one example of such a solution, DLHub by Chard et al. (2019). DLHub focuses on sharing and deploying models, offering not only deployment functionality, but also a model repository that can be used to verify results and reuse existing models. Such a model repository could enable others to benefit from the work one has done developing an AI model, as others could use the model for their own projects. Currently there is no standard method for publishing AI models, resulting in models being published in a variety of locations from GitHub to private websites (Chard et al., 2019). Therefore, creating a common location for AI resources might make the process of finding relevant resources easier.

Another proposal with a similar fundamental idea is suggested by Montes and Goertzel (2019), who argues for an AI marketplace relying on Distributed Ledger Technology (DLT). The purpose of using DLT would be to enable the decentralisation of this model repository, instead of relying on a traditional centralised structure with one company or institution who can control the repository. The authors go on to suggest that such a repository, with live deployed models, could enable AI pipelines by using multiple AI models in a network to achieve complex behaviours, where a model does further processing on the output of the previous model.

The publishing of developed resources and artefacts is also motivated by several major Natural Language Processing (NLP) conferences that are encouraging authors to ensure reproducibility of their submitted works (Moreau et al., 2019). For many projects the artefacts produced are considered part of the outcome of the project, further encouraging reproducibility through shared data and code (Moreau et al., 2019). Moreover, Moreau et al. (2019) proposes the use of workshops and journals with publications that focus on reproduction of results, maintenance and improvement of research tools and documentation to incentivise researchers to spend more time improving and maintaining a data and software foundation for future research.

4.4.8 Education

There a wide support in the literature for a more extensive offer in AI education to increase AI literacy (Allen et al., 2019; Belinchon et al., 2019; Floridi et al., 2018; Kobayashi et al., 2019; Lauterbach, 2019; Moreau et al., 2019; Robinson, 2020; Van Horn et al., 2017; Vazhayil et al., 2019; Ylipulli and Luusua, 2019). While there is information about AI available online, such as the initiatives Elements of AI[24], AI4All[25], AI Saturdays[26], BD2K Training Coodinating Center[27] and online education marketplaces that have started offering courses in AI, which may be a good place to start for those who need an introduction into the field (Luce, 2019), it may be more challenging to actually get practical experience (Allen et al., 2019). Possible ways of expanding the offer is through education in school (Belinchon et al., 2019; Floridi et al., 2018; Lauterbach, 2019; Pavao et al., 2019; Vartiainen et al., 2020; Vazhayil et al., 2019), experimentation in libraries

[24]Elements of AI https://www.elementsofai.com
[25]AI4All https://ai-4-all.org/
[26]AI Saturdays https://saturdays.ai/
[27]BD2K Training Coodinating Center http://bigdatau.org/

(Finley, 2019; Ylipulli and Luusua, 2019), tinkering in maker spaces or specialised centers (Van Horn et al., 2017). The goal is not necessarily that everyone should know the intricate details of AI design and development, but at least ensuring that people have an understanding of how the technology works and what it is capable of so they are able to make informed decisions regarding it in the future (Ienca, 2019). Further, this would be a step towards preparing for potential shifts in the job market, where some jobs lost to automation may be replaced with positions that are more tech related (Lauterbach, 2019). It is also important to include courses in the ethics of AI in the AI education (Floridi et al., 2018; Shah, 2018) to promote a future where AI researchers and developers are even more mindful and aware of potential consequences of their AI.

Research challenges for the openness of AI includes:

- Research the true impact of open AI, weighing risks of misuse against the benefits of a wider group of people gaining access to the technology (Smith and Neupane, 2018).
- Researching how AI applications can be successfully scaled up to benefit as many people as possible, including how to expand a system across cultural boundaries without creating negative bias (Smith and Neupane, 2018).
- Investigate how knowledge about AI can most effectively be shared to enable people to make decisions related to AI, both in the general public and government (Smith and Neupane, 2018).
- Explore possible solutions to the challenge that many educators do not have a background in AI.
- Investigate the impact of a race for AI technology on the willingness to accept risk, or cut corners in risk assessment or deliberative processes, and whether openness and international collaboration can reduce these risks.
- Explore how to most effectively enable a wider group of people to design, develop and use AI locally, including in the Global South (Smith and Neupane, 2018). This challenge includes both access to knowledge and training, but also necessary resources, both digital and physical.

4.5 Inclusiveness

The primary challenge of inclusiveness is ensuring that the interests of relevant social groups are considered, such that they are not treated unfairly in the application of AI systems. This is not limited to the development of AI applications for public use, but also includes AI research and innovation. An approach to promote this consideration is by enabling participation and collaboration between public, private, academia and community in the development, use and governance of AI (Floridi et al., 2018; Robinson, 2020; Yigitcanlar and Cugurullo, 2020). Further, ensuring diversity within the

technology profession (Shah, 2018), promoting teams with diverse backgrounds (Belinchon et al., 2019), both in culture, heritage and education. And adhering to ethical standards established for the relevant context, for example the European Commission's ethical guidelines for AI development (Yigitcanlar and Cugurullo, 2020).

4.5.1 Fairness

The first step towards ensuring that no group is being treated unfairly should be defining what fairness means. However, this turns out to be a deceptively non-trivial and context dependent task, as there is a great number of ways of interpreting the term and many of them are trade-offs of one another (Beretta et al., 2019). Further, these different interpretations of the term align with different interpretations of equality in democracy (Beretta et al., 2019), as discussed in section 3. Beretta et al. attempts to address this challenge in their paper *The Invisible Power of Fairness* (2019), going through various possible interpretations of the term in the context of machine learning and explaining them both in mathematical and natural language terms.

Timan and Grommé (2020) developed a framework identifying a number of challenges to fairness related to automated decision making. The framework focuses on the data processing workflow and notes potential challenges for fairness in each step, both for the intermediary results and the processes for getting to those results. Finally, the framework highlights some potential solution spaces for each step in the workflow.

Wong (2020) adapted a framework called Accountability for Reasonableness, originally developed by Daniels and Sabin (1997) for healthcare contexts, for determining whether an appropriate process has been followed to ensure algorithmic fairness. The framework consists of four conditions, which must all be met for the process to be considered fair. The four conditions, directly quoted, are as follows (Wong, 2020):

1. Publicity condition: Decisions that establish priorities in meeting [algorithmic fairness] and their rationales must be publicly accessible.

2. Relevance condition: The rationales for priority-setting decisions should aim to provide a reasonable explanation of why the priorities selected are thought the best way to progressively realize [the value the algorithm aims to provide] or the best way to meet [claims] of the defined population under reasonable [. . .] constraints. Specifically, a rationale will be "reasonable" if it appeals to evidence, reasons, and principles that are accepted as relevant by ("fair minded") people who are disposed to finding mutually justifiable terms of cooperation. An obvious device for testing the relevance of reasons is to include a broad range of stakeholders affected by these decisions so that the deliberation considers the full range of considerations people think are relevant to setting priorities.

3. Revision and appeals condition: There must be mechanisms for challenge and dispute resolution regarding priority-setting decisions and,

more broadly, opportunities for revision and improvement of policies in light of new evidence or arguments.

4. Regulative condition: There is public regulation of the process to ensure that conditions (1)–(3) are met.

4.5.2 Stakeholders

When deciding who to include in the various processes, one first has to identify who the relevant stakeholders are for an application, which may be less trivial than it sounds. Defining too sharp boundaries between groups may risk accidental exclusion, while having too vague boundaries may render the groups meaningless (Belinchon et al., 2019). Belinchon et al. (2019) argues for a global standard process for identifying stakeholders, which could help developers correctly identifying whom they need to include in the development and use processes. Such a standardised process could also make it easier for a development team to justify their choices, for example in the case of an audit.

4.5.3 Participation

Open-source development enables more interaction between developers and users in addition to allowing users to participate in the planning and development of the software, even without programming experience (Moreau et al., 2019). The general public could, for example, contribute by writing and keeping documentation up-to-date, organise and coordinate discussions on ethics, work on promoting the project to gain interest from other possible participants or even labelling data. Including domain experts in the development of AI projects could, for example, ensure that the AI operates as it should and that it is safe to deploy for the planned use case (Allen et al., 2019).

Further, very few research projects are truly open up for public participation (Moreau et al., 2019) and several papers propose to change this (Buckingham Shum et al., 2012; Moreau et al., 2019). However, while researchers may be rewarded for their efforts through publications and academic recognition resulting from the projects they contribute to, the same may not apply to members of the general public. It is therefore important to find ways to incentivise the community such that they feel rewarded for their efforts instead of being exploited (Moreau et al., 2019). However, one issue with rewarding the general public for providing or labelling data, is worker exploitation (Baird and Schuller, 2020). This has been a common critique of solutions such as Amazon's Mechanical Turk (Hara et al., 2017; Semuels, 2018). Moreau et al. (2019) suggest that enabling potential contributors to see the extent of the contributions by others in the community may motivate them to get involved. Small to Medium Enterprises (SME) may be motivated to contribute if the resulting tools or technology can be of use to them (Moreau et al., 2019). While larger companies may be motivated by external innovation from open development, the possibility of their technology being considered as a standard due to the wide adoption or support of the open source community or additional sales of related products

or services (West and Gallagher, 2006).

A challenge with attracting attention and participants to a project is the lack of visibility (Moreau et al., 2019). If those who are potentially interested in the project are not aware of its existence or there is no easy way of finding the project, they might never come across it. It must also be apparent that the project is open to contributions (Moreau et al., 2019), otherwise potential participants may simply go on with their day instead of offering their help. GitHub is widely used for open source projects, perhaps one can draw inspiration from their solution in creating a platform for discovering potentially interesting AI projects.

While the goal of inclusion is collaboration in the realisation of projects that will make a positive impact on society, there are bound to be disagreements and conflicts. Many of these disagreements will be founded on differing principals and visions of what the future should look like, and such disagreements are the core purpose of inclusion in the democratisation of AI. But there will also be disagreements founded on economic imperatives, political partisanship, false information and, if the project is openly available online, just for the sake of causing conflict. It is therefore of paramount importance to design any system in such a way that it promotes healthy and constructive disagreements, while also avoiding destructive conflicts (Buckingham Shum et al., 2012; Dobbe et al., 2020). If a fundamental conflict regarding a project cannot be resolved or appropriately addressed, one should consider the option of not proceeding with the project (Dobbe et al., 2020). Umbrello (2019) proposes the use of Value Sensitive Design to evaluate what values are embedded in decisions and to help find common ground between stakeholders with disparate interests in development of AI.

Dobbe et al. (2020) propose a set of of commitments and questions that AI projects should consider and answer in the process of designing, developing and using AI. The questions highlight challenges that projects may face and that can help guide the work with the general public.

The involvement of the general public in the development and use of AI requires development of novel techniques and methods of enabling this. Lock et al. (2020) write about the participatory methodology they developed for involving planners and policy makers in the development of a technique for including machine learning in a Planning Support System (PSS) for urban development, as well as their experience and results from going through this process. This inclusiveness on a meta level, not only in the development and use of AI, but in the development of methods to enable more people to participate in the development and use of AI, could help uncover potential solutions not apparent to experts in AI.

Buckingham Shum et al. (2012) provides a framework to classify the level of public participation in a research project, separating the depth of public involvement into four levels. I have adapted this framework to more generally describe participation in an AI project, shown in figure 2.

The first, and most basic, level includes projects where the general public does not serve any active role beyond assisting the project with obtaining the fundamental resources necessary for the development process. This can include providing data (Baird and Schuller, 2020), for example by uploading data recorded from them playing a game, or loaning compute resources such as the BOINC (Buckingham Shum et al., 2012) and Folding@Home

- **Level 4 - Collaboratory development**
  - General public participates in the development, deployment and maintenance of the project.
- **Level 3 - Participatory development**
  - General public participates in the problem definition, planning and other tasks related to the project such as promotion or coordinating participants from various stakeholder groups.
- **Level 2 - Distributed intelligence**
  - General public and/or advocacy groups participates in discussions regarding ethical issues and other challenges, and/or the labelling of data.
- **Level 1 - Crowdsourcing**
  - General public participates by loaning compute time from their machines or by providing data.

Figure 2: A framework for determining the level of public involvement in an AI project, adapted from work by Buckingham Shum et al. (2012).

projects as discussed in section 4.4.5.

The second level gives the public a more active role, where they can participate in discussions, decisions and votes about ethical issues and other challenges for the project as well as in the labelling of data (Allen et al., 2019; Buckingham Shum et al., 2012; Djeffal, 2020). Participants may also receive training in order to participate more effectively or take a test to check their ability to participate (Buckingham Shum et al., 2012). Training and tests may be particularly relevant for participants contributing to the labelling of data. Djeffal (2020) proposes randomly selecting citizens, which seems to be good on a principal level, as the selected group would most likely reflect the major opinions of the population, one might risk an under representation of minority groups or they might not be represented at all. This concern shares many similarities to algorithmic bias resulting from biased datasets, it is easy to sample data from common groups but smaller groups may not be sampled at all. To address this, one should ensure that the minority groups that have the most to lose from a decision should be sufficiently represented so they can defend their interests.

The third level involves the general public more in-depth in the process, such as data gathering and planning (Allen et al., 2019; Buckingham Shum et al., 2012). The external participants should also have, at least to some extent, shaped the problem or goal of the project (Buckingham Shum et al., 2012).

On the fourth level, the general public is completely integrated in the project developing software, setting goals, organising discussions, planning milestones and publishing papers where applicable (Allen et al., 2019; Buck-

ingham Shum et al., 2012; Musikanski et al., 2020).

It is important that the involvement of the relevant social groups in the process of ethics analysis is not just as a step in the process of development but a continuous discussion throughout the project (Moreau et al., 2019). This is an important difference as continuous involvement can identify potential issues with decisions made along the way in the project, not only address hypothetical issues at the beginning of the project or before the project is deployed. It would be very important to monitor this process closely, however, as to avoid situations where the concerns of some groups are not appreciated to their fullest extent or some social groups are not fairly represented in the discussions. Thinyane and Sassetti (2020) highlights this issue, pointing out that naive use of participatory approaches can reinforce the existing power structures and biases that already exist rather than minimising them (2020).

Musikanski et al. (2020) propose community-in-the-loop AI, where the community in its entirety is involved in the supervision of AI. While this is a step in the right direction, and it certainly could be a part of the involvement of the community, I would argue that the community would serve a quite passive role if they are not further involved in other ways.

Serious gaming is a possible tool for the generation or labelling of data (Buckingham Shum et al., 2012; Moreau et al., 2019). It is an approach where one develops a game tailored to help solve a problem or generate data by using incentives and deterrence within a game to provoke decision making and balancing of priorities (Buckingham Shum et al., 2012). A famous example is the Foldit project[28] where players solve puzzles to help solve protein folding problems.

4.5.4 Communication

Another challenge to the inclusion of the general public in the discussion of challenges, ethical or otherwise, is effectively communicating complex problems, such as complex relations, plots and numbers which can be hard to interpret or relate to without a background in the relevant domain. Storytelling is a tool that can help in addressing this challenge by constructing a narrative to create a context that illustrates the problem in question and any relevant features, using the data and plots as part of the narrative (Buckingham Shum et al., 2012).

Boundary objects and infrastructures are essentially concepts and networks of concepts that enable two or more groups to work together without having a full consensus about the meaning of the concept or construct (Buckingham Shum et al., 2012). The purpose of a boundary object is to enable the groups to coordinate, but disagreements caused by a boundary object may require discussions between the groups to realign the understanding of the boundary object to the precision necessary for work to continue (Buckingham Shum et al., 2012). These objects and infrastructures will play an important role in the inclusion across fields and domains for effective communication and collaboration (Buckingham Shum et al., 2012).

[28] http://fold.it/

4.5.5 Beneficial AI

Land and Aronson (2020) highlights a challenge for the global inclusion of AI, as AI is being used more and more in the Global South (Belinchon et al., 2019), it is important to ensure that marginalised people are not exploited for their data or underpaid labour. The goal of increasing access to AI in the Global South has to be creating value for those who needs it the most, not further increasing the wealth of large companies or institutions. Further, equality in the development and use of AI implies that AI should be deployed and used in a way that benefits those who are less able to participate, for example the Global South, such that they are able to participate on a more equal footing.

One of the major challenges towards the development of beneficial AI is a lack of incentives (Floridi et al., 2018), as the current market incentivises companies to focus on profit rather than global beneficial impact (Ghallab, 2019). Belinchon et al. (2019) proposes the creation of a beneficial AI investment fund, where successful AI companies are required to donate a percentage of their earnings to the fund, which would then distribute the money to startups or initiatives that attempt to realise the Sustainable Development Goals with the use of AI. Similarly, Montes and Goertzel (2019) proposes that a percentage of the tokens in the SingularityNet platform can be kept in a beneficial reserve, and over time get distributed to projects and external organisations considered beneficial as a way to incentivise beneficial initiatives.

The literature highlights a number of research challenges:

- Develop an international standardised process for identifying relevant stakeholder groups for different AI projects.
- Moving from more traditional eParticipation portals to interoperable services able to reach users through a variety of channels with the goal of making it simpler for users to participate (Buckingham Shum et al., 2012).
- Development of NLP tools to extract opinions and knowledge from natural language to enable the automatic summarisation of opinions, perspectives and arguments to help create overview of larger discussions (Buckingham Shum et al., 2012).
- Create an overview of AI use and policies in the Global South (Smith and Neupane, 2018) and identify potential for enabling the Global South to benefit from AI to a greater extent.
- Explore how AI innovations impacts existing social structures, what groups are at risk of being harmed and what steps can be taken to ensure the innovations contribute to a better society as a whole (Smith and Neupane, 2018).
- Research novel models for participatory design and development of AI (Smith and Neupane, 2018).

5 Community Coordinated Artificial Intelligence: A framework for the democratisation of AI

In this chapter I will be establishing my proposal for a socio-technical framework for the democratisation of AI, tackling the final research objective: *Develop a socio-technical framework for the democratisation of AI*. The purpose of the framework is to provide a set of components that can be implemented as part of an AI project or platform, with the goal of making the project or platform more democratic. Further, the framework is purposefully general, avoiding explicitly describing how something should be implemented, as it is intended to be used for projects and platforms of varying scopes. While the resulting framework can help by creating an overview of the various challenges and highlight a number of solutions for tackling them, I will emphasise that it should be considered as a starting point for further work, not as a be-all and end-all conclusion regarding what such a framework should be and do. I have named the framework Community Coordinated Artificial Intelligence (CoCoAI), emphasising the idea of tackling AI as an open community, to highlight the importance of being inclusive and spreading the control among the community of AI stakeholders rather than centralising it among a technical elite.

As discussed in section 3, the democratisation of AI is a process that involves three layers; technical, social and political. However, currently most frameworks (for instance Ahmed et al., 2020; Gao et al., 2018; Montes and Goertzel, 2019) focus on the technical side of the democratisation of AI, whereas my proposal aims to establish a relation between all three aspects.

CoCoAI is intended to be used as a tool in the process of planning new AI projects or platforms, providing a structured overview of various components that can be used to address challenges associated with the democratisation of AI, as well as pointing the reader to concrete solutions from the literature.

5.1 Components

To address the challenges discussed in chapter 4, CoCoAI contains a number of components detailed below. A number of the components will be tagged with a score describing the level of public involvement enabled by the component. The scores are based on my framework for determining the level of public involvement for an AI project (see figure 2), and range from level one to level four. It is important to note that the level of public involvement does not strictly correlate with a level of inclusiveness, as the framework concerns itself with to what extent people are able to participate, not necessarily that all interests and concerns are represented among those participating. However, as a wider community is able to participate, one can argue that there is a higher chance that more perspectives are represented.

By proposing to use these components, I do not necessarily mean that one should implement them from scratch. In many cases it would be enough

to integrate with available services or solutions. For example, one might choose to develop a crowd-sourced dataset labelling component from scratch, but simply use GitHub as an open-source AI algorithm repository.

When implementing the CoCoAI framework, one can include or leave out certain components depending on what the focus of the implementation is. There are also ways to partially implement components, which may be useful if one wishes to offer certain functionalities, but a full implementation would be too costly or not considered useful in that specific case. In the sections below, there will be examples of partial implementations of components, and later in the paper I will describe a partial implementation of the CoCoAI framework, called CoCoAI4Privacy, focused on privacy agreements.

5.1.1 Open documentation

Level of public involvement: 3
Goals:

- Make it easier for others to reuse existing resources.
- Enable the sharing of lessons learned.
- Increase transparency regarding:
  - The robustness of the AI model.
  - Steps taken to mitigate negative bias.
  - The selection of stakeholder groups.
  - The factors considered and why a decision was made during deliberation.

By keeping thorough, up-to-date documentation about the AI algorithm and the software surrounding it, it may be easier for others to integrate with, and reuse the solution. In particular, it is important to thoroughly document non-standard interfaces, file formats and so forth. Further, datasets should be documented to the extent that there is no ambiguity regarding the meaning of the data or its context (Choi and Tausczik, 2017). In addition, the development and deliberation processes should be documented such that others can learn from the experiences gained throughout the project. New projects may take this additional knowledge into consideration during planning or use it to discover new perspectives during deliberation. Further, information regarding the robustness of the AI model should also be documented and displayed prominently (Baird and Schuller, 2020; Floridi et al., 2018), so potential users of the model can evaluate whether it is suitable for their particular needs. This robustness information could potentially be displayed through the use of ML labels (Belinchon et al., 2019; Shah, 2018), a simplified visualisation of the various information to help users get a quick overview, rather than requiring them to read through large documents. In addition, a project's website could be considered part of the documentation, creating awareness about the project, the goals and how a potential participant may get involved. Relevant section: 4.3.2 on transparency.

Open documentation may enable public involvement of level three, as community members could contribute to the project by writing documentation or keeping it up to date.

5.1.2 Open standards

Level of public involvement: Not applicable
Goals:

- Discourage monopolisation of AI resources.
- Make it easier to reuse or extend existing resources.
- Reduce the complexity of integrating with available services.

By using open standards wherever applicable, it will be easier for developers to move their artefacts between services and frameworks, which reduces the chance of monopolisation (Belinchon et al., 2019; Braiek et al., 2018; Miikkulainen et al., 2019). Moreover, setting up integrations with services and frameworks may also be easier, as there may be more functionality provided through community developed software packages and libraries for popular interface standards. Relevant section: 4.1.2 on interoperability.

The use of open standards does not seem to enable any particular type of public involvement.

5.1.3 Democratic governance platform

Level of public involvement: 3
Goals:

- Decentralise the control of AI projects.
- Increase diversity in the choice of AI projects.

To decentralise the control of AI projects one should adapt a democratic form of governance. Unfortunately, the literature did not provide examples of democratic governance models for AI projects. However, one may be able to draw inspiration from democratic governance models in other contexts, such as participatory and representative democracies. Through the use of a governance platform, community members could voice problems they wish to solve or projects they are interested in creating and the community could engage in a dialogue about the problem and whether they should attempt to tackle it. Relevant section: 4.1.3 on democratic governance.

A democratic governance platform enables public involvement of level three, as community members participate in governing the project, platform and relevant resources.

5.1.4 Deliberative platform

Level of public involvement: 2

Goals:

- Enable more people to participate in debates surrounding ethics and expectations of AI.
- Increase the diversity of viewpoints brought up in relevant discussions.
- Increase the potential to unveil ethical issues or problematic areas for AI in development.
- Increase knowledge sharing regarding AI.

A deliberative platform could serve as a place to discuss the norms and expectations regarding what AI should be developed and how it should be used. The purpose is to weigh opposing opinions and interests with a diverse set of stakeholders as it may reveal additional perspectives or considerations that were not originally considered (Floridi et al., 2018; Robinson, 2020; Yigitcanlar and Cugurullo, 2020). These considerations should include how the resulting AI could be misused for malicious purposes, and whether the resulting concerns outweigh the potential benefit of the project (Ienca, 2019; Miikkulainen et al., 2019; Sudmann, 2020). Relevant sections: 4.5 on inclusiveness, 4.2.1 on ethical principles and 4.4.1 on dual-use)

The language during deliberations should be kept simple enough for non-experts to participate, so narrative based communication can be a useful tool for communicating complex topics (Buckingham Shum et al., 2012). Through participation the non-experts may gain insights into AI as they see how the project is executed and interact with people who may be more knowledgeable in AI, and thus it may be also be a useful tool for knowledge sharing. Relevant section: 4.5.4 on communication.

A democratic governance platform enables public involvement of level two, as community members participate in the deliberation regarding decisions during the project.

5.1.5 Algorithmic transparency

Level of public involvement: Not applicable
Goals:

- Make it possible for stakeholders to get explanations for why a certain decision was made about them
- Ensure decisions made by critical systems are explainable
- Enable further checking of internal AI model logic for negative bias

Technology that enables the explanation of AI decisions can help stakeholders understand how a certain decision regarding them was reached (Gao et al., 2018; Yigitcanlar and Cugurullo, 2020). One could, for example, imagine a user getting a button to request explanation along with the answer to a decision. Pressing the button could produce a short natural language explanation of the various factors and how they impacted the decision. These explanations could potentially reveal cases where there were

discriminatory factors or decisions as a step in the process of reaching a conclusion, and thus indicate a negative bias in the system (Hohman et al., 2019). This increased insight into the logic behind the system may increase a users confidence in the system, as it is less of a black box. Relevant section: 4.3.1 on algorithmic transparency.

The inclusion of methods for enabling algorithmic transparency does not seem to facilitate any particular form of public participation.

5.1.6 Open-source AI code/model repository

Level of public involvement: 4
Goals:

- Increase trust and transparency surrounding how AI is used and what types of AI are developed.
- Enable more people to make use of existing algorithms.
- Enable more people to learn from existing algorithms.
- Enable wider participation in the development.
- Increase reusability of existing models and algorithms.

In order to increase trust and transparency the algorithms developed should be open-source, along with documentation regarding how to train and use the algorithm, and what data to use. The repositories could be hosted on open-source code management systems like GitHub. Having the source code available would enable insight into the inner workings of the algorithms, so others could learn from and reuse the implementation. Furthermore, it would inform people about what data is used as input and what information is extracted, showing whether the project is extracting information that is appropriate or not. In addition, it would enable others in the community could contribute to the project, for example through the use of pull requests. By publishing the final model, other developers may use it for their projects without having to train their own model from scratch. In addition, when it comes to research projects, publishing the full algorithm and model makes it easier for other researchers to replicate results (Moreau et al., 2019). Relevant sections: 4.4.7 on AI access, 4.4.1 on dual-use and 4.1.4 on governing open resources.

Open-source AI code repository may enable public involvement of level four, as the community may participate in the development of the AI or the software surrounding it.

5.1.7 Open-source data processing

Level of public involvement: 4
Goals:

- Enable more people to participate in the development of data preprocessing pipelines.
- Enable sharing of solutions.

- Reduce the amount of work that is duplicated across projects.

A significant portion of the work when developing AI goes into setting up the data processing. By publishing the computed features, other AI developers may use the data without needing to repeat that work (Patel, 2020). Another approach is to publish the data processing pipeline itself. The pipeline can, for example, be implemented using sequences of high-level processing steps, such as scaling an image or removing stop words from a string, which can then be reused by someone else. Further, these high-level modules could be integrated into a graphical user interface, where non-programming users, relying on open-source modules developed by the community, can tweak parameters for each module to change their behaviour. One can also imagine integrating AI models into such a pipeline, to perform steps or decisions that may not be easily programmable. Relevant section: 4.4.6 on data preprocessing.

Open-source data processing may enable public involvement of level four, as the community may participate in the development of the data processing pipelines.

5.1.8 Auto ML

Level of public involvement: 4
Goals:

- Enable more people to develop AI.
- Make it easier to optimise hyper parameters.
- Make it easier to select algorithms and architectures.

Auto ML can simplify the process of developing AI to the extent that non-experts can contribute or even set out to tackle their own projects (Allen et al., 2019; Bagrow, 2020; Binnig et al., 2018; Masood and Hashmi, 2019). A user can provide a dataset and a parameter to optimise and then the algorithm will attempt to find the best combination of settings for the optimisation of the given parameter. By limiting the settings the Auto ML algorithm will explore, a user can use it to only test a set of hyper parameters, or to check what architecture from a set of architectures work best for a given problem. Relevant section: 4.4.5 on hardware access and Auto ML.

Auto ML may enable public involvement of level four, as the community may participate in the development of the AI model.

5.1.9 Distributed computing

4.4.5 **Level of public involvement:** 1

Goals:

- Increase the number of people with the capability to train AI models.
- Simplify the process of training AI using multiple machines.

- Enable people to contribute to a project without knowing programming.

Acquiring the hardware, setting up and managing the environments necessary for training AI is another challenge when working with AI. To address this, one can offer the ability to train AI on a server or a machine offered by the community. This will enable users who do not have access to powerful hardware to train their own algorithms. However, there seems to be quite a significant security risk to offer the ability to run unverified code on a user's computer, and thus it is imperative to ensure the necessary precautions have been taken such that this ability cannot be exploited. This component can be partially implemented by only offering access to hardware managed by the people organising the project, or only allowing the execution of code developed by a trusted team. Relevant section: 4.4.5 on hardware access and Auto ML, and 4.1.4 on governing open resources.

Distributed computing may enable public involvement of level one, as the community may participate by providing compute capacity with their devices.

5.1.10 Open AI model API

Level of public involvement: Not applicable
Goals:

- Increase the number of people and organisations with access to AI capabilities.
- Simplified process of deploying models.

To simplify the use of AI, one could provide an API to make it easier to integrate a trained AI model in a new application (Chard et al., 2019; Montes and Goertzel, 2019). This API could be an online web API or an AI model packaged into a component that could be easily implemented in a project. Benefits of a web API include users getting access to the most recent model version without any need to update the software. This way, heavy computation could be offloaded from the user's device and it would require little configuration by the developer. Whereas, with an AI software package, all computation could be done locally on the users device, which avoids transferring any user data. The publishing of finished models could be automated to enable people without experience in the deployment of digital services to publish their own AI models. Relevant section: 4.4.7 on AI access and 4.1.4 on governing open resources.

The inclusion of an open AI model API does not seem to facilitate any particular form of public participation in the project.

5.1.11 Open-source dataset repository

Level of public involvement: 1
Goals:

- Reduce the difficulty of finding data for a particular problem.

- Enable more people to make use of a dataset without having to create their own.
- Increase the longevity of datasets.
- Increase the transparency into the composition of datasets.
- Reduce the bias of datasets by enabling more people to analyse them.
- Reduce the chance of research becoming unreproducible due to datasets vanishing.
- Enable the community to contribute data to a project.

A common challenge in AI is assembling a dataset big enough to train a model for the problem of interest (Musikanski et al., 2020). To address this, one could host a repository of datasets, where people and projects can contribute data (Belinchon et al., 2019; Montes and Goertzel, 2019). Having an open dataset repository would enable continuous expansion of the data available, through the contributions from all the projects on the platform. Every project could then benefit from the data contributed by other projects, instead of spending time finding or assembling their own dataset. Having multiple projects labelling the same data for different purposes could increase the information richness of each example, making it easier to filter the available data down to a dataset suitable for a specific problem. Detecting and reducing negative biases in data is hard, but perhaps collaboration between experts across projects and disciplines could increase the chance of biases being detected. Further, perhaps richly labelled examples could make it easier to detect common biases. For example, if one project labelled samples with gender, another project could use that label to statistically check that they have an even distribution of samples from each gender. Relevant sections: 4.4.3 on open data, and section 4.1.4 on governing open resources.

An open-source dataset repository may enable public involvement of level one, as the community may participate by gathering or providing data for the dataset.

5.1.12 Data exploration and visualisation tools

Level of public involvement: 2
Goals:

- Improve understanding of a dataset's composition.
- Make it easier to detect biases in a set of data.
- Enable a wider community to explore and analyse data.

Data exploration and visualisation tools can play an important role by providing insights into one's data (Gao et al., 2018). These tools can help reveal negative biases or provide information about how data is related through analysis. Further, perhaps involving the community in the analysis of data can help reveal insights that otherwise would not have been realised, due to a diversity in backgrounds and experiences. However, there

are mistakes that can be made when using statistical tools without any prerequisite training, and thus it is important to ensure the tools provide sufficient guidance such that the user is able to avoid making mistakes (Kraska, 2018). Relevant section: 4.4.4 on data exploration and section 4.5 on inclusiveness.

Data exploration and visualisation tools may enable public involvement of level two, as the community may participate in the analysis of the project's dataset.

5.1.13 Crowdsourced data labelling

Level of public involvement: 2
Goals:

- Enable people with domain knowledge to participate in a project.
- Increase chance that someone notices that datasets are unbalanced.
- Reduce the work required to create datasets

In cases where the current dataset repository does not contain enough labelled data for a problem, one can call on the community to participate in creating or labelling the data (Moreau et al., 2019). To enable crowdsourcing of data labelling, one will need a data labelling interface and the necessary logic to ensure a high confidence in the resulting labels. There are several challenges with this, the most prominent being to ensure the resulting labels are of good quality, i.e., containing few false positives and false negatives, and that the distribution of images in the dataset contains as few negative biases as possible. A dataset labelled in great detail may reveal biases, i.e., if one wishes to train a model to distinguish between pictures of people who are facing towards or away from the camera, one can check the frequencies of other labels in the two data classes. One might expect that images that contain the label 'glasses' may be more common in one than the other, as one would be more likely to see glasses in pictures where the person is facing the camera. But if there is a great difference between the two classes in how many pictures are labelled with 'blue trousers', that is a bias one could argue would be irrelevant for the learning task, and one could adjust the dataset to even out this bias. Relevant sections: 4.4.3 on access to data and 4.5.3 on participation.

Crowdsourced data labelling enables public involvement of level 2, as community members can help gather, produce and label data.

5.1.14 Education platform

Level of public involvement: 3
Goals:

- Increase knowledge sharing.
- Raise general awareness of how AI works.
- Improve access to practical experience in AI.

To improve access to practical experience (Allen et al., 2019), one could provide an educational platform, based around practical examples, and developed together with domain experts in a variety of fields from the community. It could, for example, be based on projects developed in the community, where domain experts contribute with context around the problem and community members with pedagogic experience could help design assignments and problems of varying difficulty. The platform could leverage other modules from the CoCoAI framework, such as datasets from the dataset repository, example algorithms from the algorithm repository and compute from the distributed computing. Further, the educational module could include a fora where community members can discuss challenges they have encountered or share experiences. Perhaps the module could have a section where community members can share guides, including runnable examples. Or a postmortem section, where community members can write about things that went wrong and how they learned from the mistakes, so other community members don't have to repeat them. Relevant section: 4.4.8 on education and 4.4.1.

The inclusion of an education module may enable public participation of level 3, as community members can contribute in the creation of the learning resources.

5.2 Applications

As CoCoAI was developed with flexibility in mind, there is a variety of different ways in which to apply the framework. To illustrate some of this flexibility, I will provide some ideas for various applications. First of all, the framework can be used for small AI projects by simply making use of existing services that offer the functionality they need. As a step up, the framework can exist as a partial implementation, where a project uses some services to cover some functionalities, but implements their own components in the cases where they need something that is not readily available. In appendix B, I will briefly go through what such an implementation may look like.

Further, the framework can be implemented as an open-source reusable set of platform components with standardised interfaces, where projects can pick and mix between the components they are interested in. I find this use case particularly intriguing, as open-source projects can provide different implementations of the same components. For example, there can be an implementation of the Open-source dataset repository using cloud providers for storage, and another implementation that instead uses peer-to-peer technology, distributing the dataset storage across the community's devices.

Finally, at the other end of the scale, CoCoAI can be implemented as a global AI platform, serving as an international hub for AI project organisation. In this case, projects would not only benefit from the insights of their communities, but from people and advocacy groups across the globe. One benefit of such a platform is having a forum where the challenges, norms and ethics of new AI technologies are deliberated by a diverse set of stakeholders as the technologies are developed.

6 Conclusion

I started the thesis by explaining how the development and use of AI seems to primarily be a power in the hands of the few, whereas the impact of the technology affects the many. In particular, I highlighted how the technology enabled a severely imbalanced power dynamic between the AI's controller and the AI's subject. This issue is especially relevant in the context of the economic incentives in a surveillance capitalistic system. In this final chapter, I will demonstrate how my work contributes to tackling this important issue by going through the four research objectives accomplished as part of this thesis. For each objective, I will explain the contributions of the accomplishment to the initial problem and the current research on the democratisation of AI. Then, I will highlight some avenues of future research from the outset of my results, as well as how future work of system designers, politicians and users can further contribute to the democratisation of AI.

Fundamentally, the imbalance of power between those who control the AI and those who are subject to the AI is an issue of equality. Since AI is a technology where few people and organisations have the knowledge, resources and experience to develop sufficiently advanced AI, the control of AI is limited to small technical elites. Moreover, the two parties are unequal in their potential to influence one another (Hall, 2017; Jiang et al., 2017; Manheim and Kaplan, 2019). As AI can be used to extract information about peoples' interests, thoughts, political views and so forth (Kosinski et al., 2013), the party that controls AI knows significantly more about the other. This is of particular importance, as it can conflict with individuals' fundamental rights such as privacy and freedom (André et al., 2018; Floridi et al., 2018; Thwaite, 2019). This problem urges a response from societies, organisations and individuals to address how control, resources and knowledge can be shared. The democratisation of AI is an emerging AI paradigm attempting to address this exact issue. But that raises the question, what does democracy mean in the context of AI? To answer this question, I first need a foundation upon which I can base my answer.

6.1 Literature review on the democratisation of AI

In chapter 2 I addressed research objective 1 by performing a review of the literature on the democratisation of AI. As I was unable to find any published systematic review of the literature on this topic, the creation of a structured configurative review seems like a significant contribution to the research on the topic. Since the review followed a predetermined process, as described in the methodology chapter, it will be possible for other researchers to repeat the process in order to update or expand the review as new research is published. Further, when taking into account the fact that I processed all the search results for all but one query, and how broad the final queries were, this may indicate that a significant portion of the literature explicitly discussing the democratisation of AI was covered in the review. However, one still has to keep in mind that there may be literature that was not indexed by the search engines I used, and that some researchers may have chosen to use other terms for the democratisation of

AI that I did not cover or consider. After all, a structured literature review provides no guarantee that one will find all the relevant literature on a topic.

The fact that it was feasible to process all the search results (see table 1) speaks to how the democratisation of AI is still an emerging topic in the discussion surrounding the development, use and governance of AI. This observation was also supported by how the number of papers found in the review increased year over year (see figure 1). Additionally, new initiatives, such as Datatilsynet's Sandbox for Beneficial AI, UN's AI for Good, and EU's Horizon Europe work programme draft. The latter includes a call for proposals about research and development projects, with the title *Artificial intelligence, big data and democracy*. These examples further highlight the timeliness and relevancy of this thesis.

In regards to limitations of the literature review, there were a number of queries that could have uncovered additional research relevant for the democratisation of AI which I did not include due to the time constraints of my thesis. Further, the use of other search engines and literature databases may yield literature not indexed by the search engines I chose. Future research could explore these additional avenues to find topics that potentially was not uncovered in my structured review. Further, as the research on this topic evolves, it will be relevant to update the overview to ensure new discourses and topics are covered.

With an overview of the current literature on the democratisation of AI established, I was able to start addressing the question of what democracy means in relation to AI. In the literature review, I found that there were only two explicit definitions uncovered by my search (see Ienca, 2019; D. Wang et al., 2020). Further, there appeared to be various aspects of the topic that was not addressed by either definition. Thus, in order to provide a satisfactory answer to my question I decided to analyse the literature further, so I could establish my own unified definition.

6.2 Unified definition of the democratisation of AI

In chapter 3 I addressed research objective 2 by proposing a new definition for the democratisation of AI. To create the definition, I analysed how previous work had defined the democratisation of AI, both explicitly and implicitly. Further, I analysed an established definition of democracy by Held (2006), from the perspective of AI to firmly relate my definition of democratisation to a broadly accepted understanding of democracy. By analysing the literature from my search, I identified five central principles of democratisation which contribute to the definition by illustrating various mechanisms through which the technology is democratised.

The definition I established as a result of this work is as follows:

> The democratisation of artificial intelligence is the technical, social and political process towards accomplishing equality among stakeholders in the development, use and governance of AI through decentralised control, accountability, openness, transparency and inclusiveness.

This definition provides a broader understanding of the term than previous definitions found in the literature (see for instance Ienca, 2019; D. Wang et al., 2020), and importantly includes aspects of democratisation that were not previously covered. In particular, my definition improves upon existing definitions by identifying the three layers (social, technical and political) through which the process of democratisation occurs. It highlights equality as a goal for democratisation, and establishes in which AI related processes equality is important. Notably, these processes do not limit the democratisation of AI to merely be about group-decision making, but acknowledges that equality is also relevant for other aspects of AI, such as individuals' access to resources and technology. Finally, the definition presents five mechanisms that can serve as standards in the democratisation of AI. These principles are decentralised control, accountability, openness, transparency and inclusiveness (see definition). In the context of current research, the definition contributes by serving as a point that can be used to establish a common understanding among researchers regarding exactly what it means to democratise AI. This includes the ability to identify what topics are to be considered relevant for the process.

Limitations regarding my proposed definition relates back to the literature search and my academic background. As highlighted previously, there may have been literature not uncovered by my search that could have provided additional insights regarding how to define the democratisation of AI. Further, as my background is in computer science, other researchers with a deeper understanding of democracy and other fields relating to the topic may be able to identify ways to further improve it.

Now that I have answered the question of what democracy means in the context of AI, I am ready to turn my attention to what can be done to further this process. Thus, the next question is: what are the challenges faced by this process and what solutions do scholars propose to tackle them? This question is important as it highlights what problems can be tackled to make the development, use and governance of AI more democratic.

6.3 Overview of challenges and solutions for the democratisation of AI

In chapter 4 I addressed research objective 3 by providing an overview of the various challenges for the democratisation of AI and solutions presented by existing research. The overview is based upon my systematic literature review, whereby it identifies various topics discussed by the literature, highlight ways in which scholars seem to agree or disagree within each topic and discusses the solutions proposed.

As part of this effort, I created a categorisation of the various topics based upon the five principles identified in the definition. Further, I adapted a framework by Buckingham Shum et al. (2012) for determining the level of public involvement in research projects, to more generally describe public involvement in AI projects. My overview recognises the breadth of the challenges facing the democratisation of AI, spanning an interdisciplinary range from sociology, to international politics, and to technical research on the inner workings of neural networks.

I identified about 24 topics of challenges and solutions related to the democratisation of AI. Some challenges already have several solutions presented in literature. However, in other cases there appears to be gaps, as the literature did not include any solutions, such as for the democratic governance of AI projects (see section 4.1.3). This speaks to how this is an emerging paradigm and highlights that more research is needed. Many of the solutions and topics within the democratisation of AI are overlapping and interrelated, which indicates that there is an interplay between these solutions and challenges in practise. Further, many of the solutions presented face challenges of their own, and thus there is no singular solution in which all scholars agree. An example of this is the regulation of AI, where some scholars argue that regulation is necessary to ensure that the technology does not violate values and rights (see for instance Lyu et al., 2020; Shah, 2018). However, there are also researchers who argue that there should be little or no regulation of AI, as it may slow the innovation in the field, drive researchers to countries with less restrictions or that the regulations will be too slow or inflexible to adapt to the pace of AI development (for instance Baum, 2017; Ghallab, 2019; Nemitz, 2018). Therefore, it is of key importance that AI designers consider this interplay in searching for their solutions.

The literature review unveiled several topics of high significance, such as interoperability, governing open resources, algorithmic transparency, dual-use, access to data, Auto ML and access to hardware (see table 3 for a full overview). To give an indication as to what types of challenges and solutions are represented in the overview, I will present two examples:

- The majority of AI tools and frameworks are owned or backed by companies (Braiek et al., 2018). This can be problematic, as companies may have economic incentives to lock AI development within their ecosystem, for example, in order to make money from their use of cloud computing infrastructure (Braiek et al., 2018).
  - The use of open standards can ensure that developers have the ability to move between tools and frameworks (Belinchon et al., 2019; Miikkulainen et al., 2019).
  - There already exists legislation to prevent anti-competitive behaviour. Thus, ensuring that the legislation is sufficient to handle the cases regarding AI tools and frameworks may be enough to dissuade such behaviour (Ienca, 2019).
- There is a risk that open resources are abused, whether it is by tampering with data in a dataset or interrupting availability through denial-of-service attacks (Buckingham Shum et al., 2012).
  - Cryptographic techniques can be used to make it very hard for attackers to gain access, as they either need to break the encryption or gain access to a key (Buckingham Shum et al., 2012).
  - Obfuscation can be used to make data less valuable for attackers, as the valuable information in the data is either removed or separated in such a way that it becomes exceedingly difficult for an attacker to get it back (Buckingham Shum et al., 2012).

- Reputation techniques can be used to gate the access to open resources, where users are required to have a level of trust in order to gain access (Buckingham Shum et al., 2012).
- Sanctioning retroactively punishes users for bad behaviour on a platform, and the knowledge of the consequences may dissuade potential attackers from attempting (Buckingham Shum et al., 2012).

My effort provide researchers, regulators, experts and others with an overview of challenges, solutions and topics relevant to the democratisation of AI. It can help guide future research and serve as a template for the development of AI platforms and project designs. Further, for each major category in the overview I have listed future research objectives and questions highlighted by the literature in the category. These research objectives and questions can also serve as avenues of future work within and across the various subject areas covered by the overview.

6.4 Socio-technical framework for the democratisation of AI

As my background is in computer science, I decided to develop a socio-technical framework informed by the overview. This framework can be applied in the future development of AI projects and platforms to promote the beneficial development and use of AI. In chapter 5 I addressed research objective 4 by proposing a socio-technical framework for the democratisation of AI, called *Community Coordinated Artificial Intelligence*. This framework is enabled by the work done in the previous chapters of the thesis. However, in the process of developing this framework I experienced that a number of the challenges from the overview, while important, did not seem appropriate to tackle with a socio-technical framework. A good example of this is the regulation of AI, which clearly should rather be tackled through a regulatory framework, such as the framework recently proposed by the European Commission (see European Commission, 2021). This also informs the scope of the socio-technical framework, focusing on the micro-political processes related directly to the development, use and governance of AI, rather than processes such as those addressed by the aforementioned regulatory framework. I will argue, however, that the framework could also be applicable to macro-political non-governmental processes such as in an international AI platform.

To give an indication of the ways in which CoCoAI addresses the challenges of democratising AI, I will give a brief overview over some of the components included in the framework:

- **Open documentation** addresses the need for appropriate information regarding the AI solutions and resources available. This includes, for example, documenting datasets such that there is no ambiguity regarding what the data means (Choi and Tausczik, 2017). Further, information regarding the robustness of available AI services is necessary to evaluate if they are appropriate for a use case (Baird and Schuller, 2020; Floridi et al., 2018).

- **Deliberative platform** tackles the challenge of ensuring that the various perspectives, interests and concerns relevant for a decisions are appropriately considered when making decisions regarding the development and use of AI. By involving the wider community in the process of deliberation, one may reveal additional perspectives or insights that otherwise may not have been considered (Floridi et al., 2018; Robinson, 2020; Yigitcanlar and Cugurullo, 2020).

- **Open-source data processing** addresses the challenge of getting data from the source format to the format required by the AI algorithm, by enabling projects to reuse the solutions developed by other projects. This can, for example, be done through sharing computed AI features (Patel, 2020).

- **Auto ML** enables more people to develop AI algorithms to address their own interests, by simplifying or automating the process of selecting AI algorithms, hyper parameters and architectures (Allen et al., 2019; Bagrow, 2020; Binnig et al., 2018; Masood and Hashmi, 2019).

- **Open-source dataset repository** addresses the challenge of getting data for an AI problem (Belinchon et al., 2019; Montes and Goertzel, 2019; Musikanski et al., 2020). A dataset repository enables projects and community members to make use of existing datasets, but also to contribute data to the datasets available.

- **Crowdsourced data labelling** enables the community to contribute to the project by creating or labelling data to expand the dataset (Moreau et al., 2019). By combining crowdsourced data labelling with the open-source dataset repository, one can have multiple projects contributing to the labelling of the data.

The primary contribution of the framework is providing AI platform developers and project organisers with a selection of components they can implement to make the processes for developing, using and governing AI in the related project or platform more aligned with democratic principles. In relation to current research, the framework serves as a starting point for further development and calibration with existing practices. Future research can expand upon the framework by identifying additional components that would contribute to the democratisation of AI. Further, each component can be updated with new solutions or details as scholars investigate alternatives and identify new opportunities. Finally, a potentially fruitful avenue of work would be providing an implementation of the various components in an open-source modular fashion. This will make it easier for interested developers to compile a platform from the various components, rather than developing each component from scratch. Further, an open source implementation may enable projects to be more independent from existing proprietary services, which would allow them to maintain control over more of their platform.

6.5 Impact

Currently, much of the research on the democratisation of AI appears to be quite scattered, presenting differing understandings of what the term means and working on individual challenges without a recognition of the topic as a whole (see for instance Shang et al., 2019; Sudmann, 2020; D. Wang et al., 2020). The work in this thesis contributes an anchor point and foundation for the future of research on the democratisation of AI, enabling an understanding of the topic as a cohesive whole rather than as a set of isolated issues. This is recognised on a transnational level, for example by EU's Horizon Europe work programme draft for 2021 - 2022, in the approved call for proposals with the title: *Artificial intelligence, big data and democracy*. This shows the recognition of the importance of this topic among EU policy makers and research leaders. The objectives include introducing value-based frameworks for AI governance, as well as protecting and reinforcing values and rights in relation to AI. This will make resources available for further research into the topic of this thesis.

Beyond research, this work also enables societal efforts to promote beneficial AI by proposing CoCoAI, a framework showing how democratic processes can be used as catalysts to promote AI development, use and government that promotes the rights, values, interests and concerns recognised by the society in its complexity. Further, my work shows how the democratisation of AI has to happen on all levels in society, ranging from international politics to the individual person.

On the global scale we find issues such as a race dynamic between nation states. This is a very important issue, as a race dynamic may encourage participants to cut corners in regards to safety, ethics and regulation in order to stay ahead of the competition (Cave and ÓhÉigeartaigh, 2018). The disregard for these important considerations may lead to AI solutions that are actively detrimental to society (Cave and ÓhÉigeartaigh, 2018), for example by discriminating against minority groups or normalising the breach of fundamental rights such as privacy or autonomy. The regulatory framework for AI, recently proposed by the European Commission (European Commission, 2021), demonstrates recognition of the importance of ensuring that the development and use of AI does not happen at the expense of people's rights. The purpose of the framework is to regulate AI to avoid unacceptable risks in the implementation and use of the technology. This shows initiatives on the transnational level to regulate AI with a value-based approach. It comes in addition to EU initiatives to use regulation of competition against the dominance of tech giants. But there is a need for a more integrated approach towards the democratisation of AI on an transnational level.

On a national level, I show how the democratisation of AI depends on a well-functioning democratic state that can hold AI developers and users accountable for behaving responsibly. This presupposes a democratic process to produce regulation and incentives that align with the interest of the people. An example of the result from such a process is the Norwegian national strategy for AI[29]. This shows willingness of the government to use AI

[29]Norwegian national strategy for AI https://www.regjeringen.no/en/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/

to create value in collaboration with business interests, but highlights that it must be regulated and technical solutions must be developed that allows the use of data while respecting the privacy of individuals. Such initiatives could be further developed to also include the democratisation of AI, in line with the CoCoAI framework.

On an organisational level, we have challenges such as inclusiveness in various AI processes. Inclusiveness can help surface interests or concerns that might not otherwise have been considered by the AI's developers (Belinchon et al., 2019; Floridi et al., 2018; Robinson, 2020; Shah, 2018; Yigitcanlar and Cugurullo, 2020). By increasing the inclusiveness in the development, use and governance of AI projects, we can promote AI solutions that are beneficial for more people, and avoid solutions that discriminate or otherwise unfairly treat certain social groups. As a contribution to this challenge, I adapted a framework by Buckingham Shum et al. (2012) for rating the level of involvement of the general public in a research project, to more generally describe involvement in an AI project (see figure 2).

On the individual level, we have challenges such as education, as most people do not have any considerable understanding of how AI works or how to make decisions in relation to it. Further, there is a big demand for AI expertise, but a very low availability of it. Thus, there is a need for more comprehensive AI education offers (Allen et al., 2019; Belinchon et al., 2019; Floridi et al., 2018; Kobayashi et al., 2019; Lauterbach, 2019; Moreau et al., 2019; Robinson, 2020; Van Horn et al., 2017; Vazhayil et al., 2019; Ylipulli and Luusua, 2019), so that AI will become more accessible for less wealthy organisations and community projects. Education can also help ensure people have a basic understanding of how AI works, so they can make informed decisions regarding what services they will use in their day-to-day lives (Ienca, 2019). Today, increasing number of countries introduce programming as a compulsory school subject in primary education. An introduction to the basic concepts of AI could be a topic covered in such a subject.

A main contribution of my thesis is therefore to demonstrate the breadth of these challenges and how important they are for ensuring that AI will be a technology that is beneficial for our society in the long run, not just a good market opportunity for a few wealthy companies in the short-term. However, I also provide an overview of the various solutions proposed to tackle these challenges, so that researchers and other stakeholders can get a quick overview of potential approaches in the face of each challenge.

My new framework called Community Coordinated Artificial Intelligence, demonstrates how AI projects and platforms can use socio-technical components to ensure their AI and its use are beneficial. The framework impacts society by changing how AI is developed and used, resulting in solutions that are beneficial for society as a whole. Further, it impacts organisations by promoting open AI technology, resources and knowledge which provides opportunities for more organisations to make use of AI. Additionally, it impacts individuals in society by promoting education, knowledge sharing, transparency and beneficial solutions. This will help individuals in gaining a deeper understanding of the technology and thereby, make more informed decisions in relation to AI services they use. Additionally, they will have access to more beneficial AI solutions developed through democratic corporate

processes or as a result of more grassroots AI projects stemming from the wider accessibility of the technology, knowledge and resources required for the development.

CoCoAI focuses on the micro-political level in this view, as it is primarily relevant for encouraging local effects through AI projects and platforms. However, the framework can also be relevant for the macro-political level, since it can be implemented as an open global AI platform. At this level, deliberation and governance processes are not limited to one or a few local organisations or communities, but can include societies and advocacy groups across the globe. One can, for example, imagine a macro-political implementation that is governed by an international democratic institution, where the platform is open for people to organise their own AI projects. The platform can serve as a hub where AI projects on the platform, as well as companies with proprietary projects, can open their decisions for public deliberation, benefiting from the diversity in background and culture of the people involved. Governments could potentially use these public deliberations to guide their efforts in developing, altering or removing AI regulations.

Importantly, the framework contributes to the beneficial future of AI in a number of ways. First of all, it focuses on the involvement of people in the various AI processes. Perhaps most importantly, is the involvement in the deliberation of the various decisions to be made during the development, use and governance of the AI. But CoCoAI also promotes involving the community directly in the development of software and algorithms, as well as the surrounding tasks such as gathering and labelling data, organising AI projects, as well as producing documentation and keeping it up to date. This inclusiveness does not only contribute to how beneficial the resulting AI is, but it also helps dispersing AI knowledge in the community, as participants learn from seeing how the project is executed and by interacting with people who are knowledgeable on the topic. This knowledge sharing is also directly promoted through the education module, where projects can, for example, make assignments using the dataset of the project or write about things they learned throughout the process so others can also learn from their experience. Further, as CoCoAI promotes the use of open standards and ensuring that the solutions developed are easy to reuse, thus AI community as a whole can benefit from increased access to AI resources and technology.

Finally, I have established a contribution to the issue raised in the beginning of this thesis. I tackle the inequality of control by leaning on firmly established democratic principles. By using democratic processes in the development, use and governance of AI, projects can identify relevant rights and values that are potentially challenged or threatened by the outcome of the projects. By providing the people, who may be impacted by the decisions throughout the project, a way of influencing the decisions made, they can defend their own interests and thereby avoid outcomes that would impact them negatively. Further, subjects of AI would gain insight into how decisions about them are made and what information is extracted about them, which would somewhat reduce the imbalance of power between them and the AI's controller. I do not claim that my framework solves the initial problem in its entirety, after all the democratisation of AI is also dependent

on, amongst other things, appropriate regulation of the technology and its users. But I believe my framework can serve as a stepping stone in addressing these issues and making AI more beneficial for society as a whole.

References

Ahmed, S., Mula, R. S., & Dhavala, S. S. (2020). A Framework for Democratizing AI. *arXiv:2001.00818 [cs, stat]*. http://arxiv.org/abs/2001.00818

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Allen, B., Agarwal, S., Kalpathy-Cramer, J., & Dreyer, K. (2019). Democratizing AI. *Journal of the American College of Radiology*, *16*(7), 961–963. https://doi.org/10.1016/j.jacr.2019.04.023

André, Q., Carmon, Z., Wertenbroch, K., Crum, A., Frank, D., Goldstein, W., Huber, J., van Boven, L., Weber, B., & Yang, H. (2018). Consumer Choice and Autonomy in the Age of Artificial Intelligence and Big Data. *Customer Needs and Solutions*, *5*(1), 28–37. https://doi.org/10.1007/s40547-017-0085-8

Arogyaswamy, B. (2020). Big tech and societal sustainability: An ethical framework. *AI & SOCIETY*, *35*(4), 829–840. https://doi.org/10.1007/s00146-020-00956-6

Bagrow, J. (2020). Democratizing AI: Non-expert design of prediction tasks. *PeerJ Computer Science*, *6*. https://doi.org/10.7717/PEERJ-CS.296

Baird, A., & Schuller, B. (2020). Considerations for a More Ethical Approach to Data in AI: On Data Representation and Infrastructure [Publisher: Frontiers Media SA]. *Frontiers in Big Data*, *3*. https://doi.org/10.3389/fdata.2020.00025

Baum, S. D. (2017). On the promotion of safe and socially beneficial artificial intelligence. *AI & SOCIETY*, *32*(4), 543–551. https://doi.org/10.1007/s00146-016-0677-0

Belinchon, E., Bollmann, B., Cussins Newman, J., Jobin, A., & Nakonz, J. (2019). *Towards an Inclusive Future in AI. A Global Participatory Process* (SSRN Scholarly Paper No. ID 3505425). Social Science Research Network. Rochester, NY. Retrieved October 28, 2020, from https://papers.ssrn.com/abstract=3505425

Beretta, E., Santangelo, A., Lepri, B., Vetro, A., & De Martin, J. C. (2019). The Invisible Power of Fairness. How Machine Learning Shapes Democracy. In M. J. Meurs & F. Rudzicz (Eds.), *Advances in Artificial Intelligence* (pp. 238–250). Springer International Publishing Ag.

Binnig, C., Buratti, B., Chung, Y., Cousins, C., Kraska, T., Shang, Z., Upfal, E., Zeleznik, R., & Zgraggen, E. (2018). Towards Interactive Curation & Automatic Tuning of ML Pipelines. *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, 1–4. https://doi.org/10.1145/3209889.3209891

Bostrom, N. (2017). Strategic Implications of Openness in AI Development. *Global Policy*, *8*(2), 135–148. https://doi.org/10.1111/1758-5899.12403

Braiek, H. B., Khomh, F., & Adams, B. (2018). The Open-Closed Principle of Modern Machine Learning Frameworks. *2018 IEEE/ACM 15th*

International Conference on Mining Software Repositories (MSR), 353–363. https://ieeexplore.ieee.org/document/8595219

Bryman, A. (2016). *Social Research Methods* (5th edition). Oxford University Press.

Buckingham Shum, S., Aberer, K., Schmidt, A., Bishop, S., Lukowicz, P., Anderson, S., Charalabidis, Y., Domingue, J., de Freitas, S., Dunwell, I., Edmonds, B., Grey, F., Haklay, M., Jelasity, M., Karpištšenko, A., Kohlhammer, J., Lewis, J., Pitt, J., Sumner, R., & Helbing, D. (2012). Towards a global participatory platform: Democratising open data, complexity science and collective intelligence. *The European Physical Journal Special Topics*, *214*(1), 109–152. https://doi.org/10.1140/epjst/e2012-01690-3

Carter, D. (2020). Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role? *Business Information Review*, *37*(2), 60–68. https://doi.org/10.1177/0266382120923962

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, *8*(8), 832. https://doi.org/10.3390/electronics8080832

Cave, S., & ÓhÉigeartaigh, S. S. (2018). An AI Race for Strategic Advantage: Rhetoric and Risks. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 36–40. https://doi.org/10.1145/3278721.3278780

Chard, R., Li, Z., Chard, K., Ward, L., Babuji, Y., Woodard, A., Tuecke, S., Blaiszik, B., Franklin, M. J., & Foster, I. (2019). DLHub: Model and Data Serving for Science. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 283–292. https://doi.org/10.1109/IPDPS.2019.00038

Choi, J., & Tausczik, Y. (2017). Characteristics of Collaboration in the Emerging Practice of Open Data Analysis. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 835–846. https://doi.org/10.1145/2998181.2998265

Chui, M., Harrysson, M., Manyika, J., Roberts, R., Chung, R., Nel, P., & van Heteren, A. (2018). Applying AI for social good. Retrieved May 10, 2021, from https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good

Chung, Y., Kraska, T., Polyzotis, N., Tae, K. H., & Whang, S. E. (2019). Slice Finder: Automated Data Slicing for Model Validation. *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 1550–1553. https://doi.org/10.1109/ICDE.2019.00139

Clarke, R. (2019). Principles and business processes for responsible AI. *Computer Law & Security Review*, *35*(4), 410–422. https://doi.org/10.1016/j.clsr.2019.04.007

Couchman, H., & Lemos, A. P. (2019). Policing by Machine: Predictive policing and the threat to our rights. Retrieved October 1, 2020, from https://www.libertyhumanrights.org.uk/wp-content/uploads/2020/02/LIB-11-Predictive-Policing-Report-WEB.pdf

Daniels, N., & Sabin, J. (1997). Limits to Health Care: Fair Procedures, Democratic Deliberation, and the Legitimacy Problem for Insurers.

Philosophy & Public Affairs, *26*(4), 303–350. https://doi.org/https://doi.org/10.1111/j.1088-4963.1997.tb00082.x

Datatilsynet. (2020). Sandbox for responsible artificial intelligence. Retrieved May 6, 2021, from https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/

Djeffal, C. (2020). AI, Democracy and the Law. *The Democratization of Artificial Intelligence* (pp. 255–284). transcript Verlag.

Dobbe, R., Gilbert, T., & Mintz, Y. (2020). Hard choices in artificial intelligence: Addressing Normative Uncertainty through Sociotechnical Commitments, 242. https://doi.org/10.1145/3375627.3375861

Donahoe, E., & Metzger, M. M. (2019). Artificial Intelligence and Human Rights. *Journal of Democracy*, *30*(2), 115–126. https://doi.org/10.1353/jod.2019.0029

D'Souza, D. (2019). Tech Lobby: Internet Giants Spend Record Amounts, Electronics Firms Trim Budgets. Retrieved April 26, 2021, from https://www.investopedia.com/tech/what-are-tech-giants-lobbying-trump-era/

Elster, J. (1998). *Deliberative Democracy*. Cambridge University Press. Retrieved May 4, 2021, from https://repository.library.georgetown.edu/handle/10822/909083

Etzioni, O. (2018). Point: Should AI technology be regulated? yes, and here's how. *Communications of the ACM*, *61*(12), 30–32. https://doi.org/10.1145/3197382

European Commission. (2021). Regulatory framework on AI. Retrieved May 6, 2021, from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Parliament. (2016). GDPR. Retrieved October 2, 2020, from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

Feenberg, A. (1991). *Critical theory of technology*. Oxford University Press.

Finley, T. K. (2019). The Democratization of Artificial Intelligence: One Library's Approach. *Information Technology and Libraries*, *38*(1), 8–13. https://doi.org/10.6017/ital.v38i1.10974

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Trans. Inf. Syst.*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Friend, Z. (2013). Predictive Policing: Using Technology to Reduce Crime. Retrieved October 1, 2020, from https://leb.fbi.gov/articles/featured-articles/predictive-policing-using-technology-to-reduce-crime

Gao, J., Wang, W., Zhang, M., Chen, G., Jagadish, H. V., Li, G., Ng, T. K., Ooi, B. C., Wang, S., & Zhou, J. (2018). PANDA: Facilitating Usable AI Development. *arXiv:1804.09997 [cs]*. http://arxiv.org/abs/1804.09997

Garvey, C. (2018). A Framework for Evaluating Barriers to the Democratization of Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://ojs.aaai.org/index.php/AAAI/article/view/12194

Ghallab, M. (2019). Responsible AI: Requirements and challenges. *AI Perspectives*, *1*(1), 3. https://doi.org/10.1186/s42467-019-0003-z

Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews*, *1*(1), 28. https://doi.org/10.1186/2046-4053-1-28

Guss, W. H., Houghton, B., Topin, N., Wang, P., Codel, C., Veloso, M., & Salakhutdinov, R. (2019). MineRL: A Large-Scale Dataset of Minecraft Demonstrations. *arXiv:1907.13440 [cs, stat]*. http://arxiv.org/abs/1907.13440

Habermas, J. (2015). *Between facts and norms: Contributions to a discourse theory of law and democracy*. John Wiley & Sons.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2125–2126. https://doi.org/10.1145/2939672.2945386

Hall, S. (2017). How Artificial Intelligence Is Changing the Insurance Industry. Retrieved September 30, 2020, from https://www.naic.org/cipr_newsletter_archive/vol22_ai.pdf

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. (2017). A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. *arXiv:1712.05796 [cs]*. http://arxiv.org/abs/1712.05796

Harwell, D., & Dou, E. (2020). Huawei tested AI software that could recognize Uighur minorities and alert police, report says. *Washington Post*. Retrieved January 10, 2021, from https://www.washingtonpost.com/technology/2020/12/08/huawei-tested-ai-software-that-could-recognize-uighur-minorities-alert-police-report-says/

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1026–1034. https://doi.org/10.1109/ICCV.2015.123

Held, D. (2006). *Models of democracy*. Stanford University Press.

Hertel-Fernandez, A., Skocpol, T., & Sclar, J. (2018). When political mega-donors join forces: How the Koch network and the Democracy Alliance influence organized US politics on the right and left. *Studies in American Political Development*, *32*(2), 127–165.

Hohman, F., Kahng, M., Pienta, R., & Chau, D. H. (2019). Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics*, *25*(8), 2674–2693. https://doi.org/10.1109/TVCG.2018.2843369

Holford, W. D. (2019). The future of human creative knowledge work within the digital economy. *Futures*, *105*, 143–154. https://doi.org/10.1016/j.futures.2018.10.002

Hutt, D. B. (2018). Republicanism, Deliberative Democracy, and Equality of Access and Deliberation. *Theoria*, *84*(1), 83–111. https://doi.org/10.1111/theo.12138

Ienca, M. (2019). Democratizing cognitive technology: A proactive approach. *Ethics and Information Technology*, *21*(4), 267–280. https://doi.org/10.1007/s10676-018-9453-9

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, *51*(8), 56–59. https://doi.org/10.1109/MC.2018.3191268

Jain, N., Manikonda, L., Hernandez, A. O., Sengupta, S., & Kambhampati, S. (2018). Imagining an Engineer: On GAN-Based Data Augmentation Perpetuating Biases. *arXiv:1811.03751 [cs, stat]*. http://arxiv.org/abs/1811.03751

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4). https://doi.org/10.1136/svn-2017-000101

Johansen, J., Pedersen, T., Fischer-Hübner, S., Johansen, C., Schneider, G., Roosendaal, A., Zwingelberg, H., Sivesind, A. J., & Noll, J. (2021). Privacy Labelling and the Story of Princess Privacy and the Seven Helpers. *arXiv:2012.01813 [cs]*. http://arxiv.org/abs/2012.01813

Keohane, R. O., & Nye Jr, J. S. (1998). Power and interdependence in the information age. *Foreign Aff.*, *77*, 81.

Kobayashi, Y., Ishibashi, M., & Kobayashi, H. (2019). How will "democratization of artificial intelligence" change the future of radiologists? *Japanese Journal of Radiology*, *37*(1), 9–14. https://doi.org/10.1007/s11604-018-0793-5

Kojima, R., Ishida, S., Ohta, M., Iwata, H., Honma, T., & Okuno, Y. (2020). kGCN: A graph-based deep learning framework for chemical structures. *Journal of Cheminformatics*, *12*. https://doi.org/10.1186/s13321-020-00435-6

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Kraska, T. (2018). Northstar: An interactive data science system. *Proceedings of the VLDB Endowment*, *11*(12), 2150–2164. https://doi.org/10.14778/3229863.3240493

Kriebitz, A., & Lütge, C. (2020). Artificial Intelligence and Human Rights: A Business Ethical Assessment [Publisher: Cambridge University Press]. *Business and Human Rights Journal*, *5*(1), 84–104. https://doi.org/10.1017/bhj.2019.28

Land, M. K., & Aronson, J. D. (2020). Human Rights and Technology: New Challenges for Justice and Accountability. *Annual Review of Law and Social Science*, *16*(1), 223–240. https://doi.org/10.1146/annurev-lawsocsci-060220-081955

Lauterbach, A. (2019). Artificial intelligence and policy: Quo vadis? *Digital Policy, Regulation and Governance*, *21*(3), 238–263. https://doi.org/10.1108/DPRG-09-2018-0054

Lee, D. J.-L. (2020). Towards an Integrated Solution for Intelligent Visual Data Discovery. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3334480.3375035

Lepore, J. (2020). Scientists use big data to sway elections and predict riots — welcome to the 1960s. *Nature*, *585*(7825), 348–350. https://doi.org/10.1038/d41586-020-02607-8

Lewis, A. (2010). Metafilter: User-driven discontent. Retrieved November 13, 2019, from https://www.metafilter.com/95152/Userdriven-discontent

Liang, J., Meyerson, E., Hodjat, B., Fink, D., Mutch, K., & Miikkulainen, R. (2019). Evolutionary neural AutoML for deep learning. *Proceedings of the Genetic and Evolutionary Computation Conference*, 401–409. https://doi.org/10.1145/3321707.3321721

Lindblom, C. E. (1993). *The policy-making process* (3rd ed.). Prentice Hall.

Lock, O., Bain, M., & Pettit, C. (2020). Towards the collaborative development of machine learning techniques in planning support systems - a Sydney example. *Environment and Planning B-Urban Analytics and City Science*, 2399808320939974. https://doi.org/10.1177/2399808320939974

Luce, L. (2019). Democratization and Impacts of AI. In L. Luce (Ed.), *Artificial Intelligence for Fashion: How AI is Revolutionizing the Fashion Industry* (pp. 185–195). Apress. https://doi.org/10.1007/978-1-4842-3931-5_12

Lyu, L., Li, Y., Nandakumar, K., Yu, J., & Ma, X. (2020). How to Democratise and Protect AI: Fair and Differentially Private Decentralised Deep Learning. *IEEE Transactions on Dependable and Secure Computing*, 1–1. https://doi.org/10.1109/TDSC.2020.3006287

Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. *Omega*, *31*(2), 83–96. https://doi.org/10.1016/S0305-0483(03)00016-1

Manheim, K. M., & Kaplan, L. (2019). Artificial Intelligence: Risks to Privacy and Democracy. *Yale Journal of Law & Technology*, *21*(1), 106–188. https://papers.ssrn.com/abstract=3273016

Masood, A., & Hashmi, A. (2019). Democratization of AI Using Cognitive Services. *Cognitive Computing Recipes: Artificial Intelligence Solutions Using Microsoft Cognitive Services and TensorFlow* (pp. 1–17). Apress. https://doi.org/10.1007/978-1-4842-4106-6_1

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, *114*(48), 12714. https://doi.org/10.1073/pnas.1710966114

Miikkulainen, R., Greenstein, B., Hodjat, B., & Smith, J. (2019). Better Future through AI: Avoiding Pitfalls and Guiding AI Towards its Full Potential. *arXiv:1905.13178 [cs]*. http://arxiv.org/abs/1905.13178

Montes, G. A., & Goertzel, B. (2019). Distributed, decentralized, and democratized artificial intelligence. *Technological Forecasting and Social Change*, *141*, 354–358. https://doi.org/10.1016/j.techfore.2018.11.010

Moreau, E., Vogel, C., & Barry, M. (2019). A Paradigm for Democratizing Artificial Intelligence Research. In A. Esposito, A. M. Esposito, & L. C. Jain (Eds.), *Innovations in Big Data Mining and Embedded Knowledge* (pp. 137–166). Springer International Publishing. https://doi.org/10.1007/978-3-030-15939-9_8

Mozur, P. (2018). Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras. *The New York Times*. Retrieved October 1, 2020, from https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html

Muehlhauser, L., & Salamon, A. (2012). Intelligence Explosion: Evidence and Import. In A. H. Eden, J. H. Moor, J. H. Søraker, & E. Steinhart (Eds.), *Singularity Hypotheses: A Scientific and Philosophical Assessment* (pp. 15–42). Springer. https://doi.org/10.1007/978-3-642-32560-1_2

Musikanski, L., Rakova, B., Bradbury, J., Phillips, R., & Manson, M. (2020). Artificial Intelligence and Community Well-being: A Proposal for an Emerging Area of Research. *International Journal of Community Well-Being*, *3*(1), 39–55. https://doi.org/10.1007/s42413-019-00054-6

Nakamura, M., & Yamakawa, H. (2016). A Game-Engine-Based Learning Environment Framework for Artificial General Intelligence. In A. Hirose, S. Ozawa, K. Doya, K. Ikeda, M. Lee, & D. Liu (Eds.), *Neural Information Processing* (pp. 351–356). Springer International Publishing. https://doi.org/10.1007/978-3-319-46687-3_39

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133). https://doi.org/10.1098/rsta.2018.0089

Neuman, M., Bitton, A., & Glantz, S. (2002). Tobacco industry strategies for influencing European Community tobacco advertising legislation. *The Lancet*, *359*(9314), 1323–1330. https://doi.org/10.1016/S0140-6736(02)08275-2

O'Hare, R. (2017). Research collaboration aims to improve breast cancer diagnosis using AI. Retrieved May 10, 2021, from https://www.imperial.ac.uk/news/183293/research-collaboration-aims-improve-breast-cancer/

O'Leary, Z. (2017). *The Essential Guide to Doing Your Research Project* (Third edition). SAGE Publications Ltd.

Olson, R. S., Sipper, M., Cava, W. L., Tartarone, S., Vitale, S., Fu, W., Orzechowski, P., Urbanowicz, R. J., Holmes, J. H., & Moore, J. H. (2018). A System for Accessible Artificial Intelligence. In W. Banzhaf, R. S. Olson, W. Tozier, & R. Riolo (Eds.), *Genetic Programming Theory and Practice XV* (pp. 121–134). Springer International Publishing. https://doi.org/10.1007/978-3-319-90512-9_8

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

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

O'Sullivan, A., & Thierer, A. (2018). Counterpoint: Regulators should allow the greatest space for AI innovation. *Communications of the ACM*, *61*(12), 33–35. https://doi.org/10.1145/3241035

Patel, J. (2020). The Democratization of Machine Learning Features, 136–141. https://doi.org/10.1109/IRI49571.2020.00027

Pavao, A., Kalainathan, D., Sun-Hosoya, L., Bennett, K., & Guyon, I. (2019). Design and Analysis of Experiments: A Challenge Approach in Teaching. https://hal.inria.fr/hal-02415639

Peters, M. A. (2019). Platform ontologies, the AI crisis and the ability to hack humans 'An algorithm knows me better than I know myself'. *Educational Philosophy and Theory*, *0*(0), 1–9. https://doi.org/10.1080/00131857.2019.1618227

Raso, F. A., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018). *Artificial Intelligence & Human Rights: Opportunities & Risks* (SSRN Scholarly Paper No. ID 3259344). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3259344

Robinson, S. C. (2020). Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). *Technology in Society*, 101421. https://doi.org/10.1016/j.techsoc.2020.101421

Schalkoff, R. J. (1990). *Artificial intelligence: An engineering approach*. McGraw-Hill New York, NY, USA.

Schneider, I. (2020). Democratic Governance of Digital Platforms and Artificial Intelligence? : Exploring Governance Models of China, the US, the EU and Mexico. *JeDEM - eJournal of eDemocracy and Open Government*, *12*(1), 1–24. https://doi.org/10.29379/jedem.v12i1.604

Schneier, B. (2015). *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. W. W. Norton & Company.

Seale, C. (Ed.). (2018). *Researching Society and Culture* (Fourth edition). SAGE Publications Ltd.

Semuels, A. (2018). The Internet Is Enabling a New Kind of Poorly Paid Hell. Retrieved April 29, 2021, from https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/

Shah, H. (2018). Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128). https://doi.org/10.1098/rsta.2017.0362

Shang, Z., Zgraggen, E., Buratti, B., Kossmann, F., Eichmann, P., Chung, Y., Binnig, C., Upfal, E., & Kraska, T. (2019). Democratizing Data Science through Interactive Curation of ML Pipelines. *Proceedings of the 2019 International Conference on Management of Data*, 1171–1188. https://doi.org/10.1145/3299869.3319863

Smith, M., & Neupane, S. (2018). Artificial intelligence and human development: Toward a research agenda. Retrieved October 28, 2020, from https://idl-bnc-idrc.dspacedirect.org/handle/10625/56949

Sudmann, A. (2018). On the Media-political Dimension of Artificial Intelligence: Deep Learning as a Black Box and OpenAI. *Digital Culture & Society*, *4*(1), 181–200. https://doi.org/10.14361/dcs-2018-0111

Sudmann, A. (2020). The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms. *The Democratization of Artificial Intelligence* (pp. 9–32). transcript Verlag.

Thinyane, H., & Sassetti, F. (2020). Towards a Human Rights-Based Approach to AI: Case Study of Apprise. In D. R. Junio & C. Koopman (Eds.), *Evolving Perspectives on ICTs in Global Souths* (pp. 33–47). Springer International Publishing. https://doi.org/10.1007/978-3-030-52014-4_3

Thwaite, A. (2019). Literature review on elections, political campaigning and democracy. Retrieved November 14, 2019, from https://comprop.oii . ox . ac . uk / wp - content / uploads / sites / 93 / 2019 / 09 / OxTEC - Literature-Review-Alice-Thwaite-Report-25-09-19.pdf

Timan, T., & Grommé, F. (2020). *A framework for social fairness; Insights from two algorithmic decision-making controversies in The Netherlands*.

Umbrello, S. (2019). Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach. *Big Data and Cognitive Computing*, *3*(1), 5. https://doi.org/10.3390/bdcc3010005

Van Horn, J. D., Fierro, L., Kamdar, J., Gordon, J., Stewart, C., Bhattrai, A., Abe, S., Lei, X., O'Driscoll, C., Sinha, A., Jain, P., Burns, G., Lerman, K., & Ambite, J. L. (2017). Democratizing data science through data science training. *Biocomputing 2018* (pp. 292–303). World Scientific. https://doi.org/10.1142/9789813235533_0027

Vartiainen, H., Tedre, M., & Valtonen, T. (2020). Learning machine learning with very young children: Who is teaching whom? *International Journal of Child-Computer Interaction*, *25*, 100182. https://doi.org/10.1016/j.ijcci.2020.100182

Vazhayil, A., Shetty, R., Bhavani, R. R., & Akshay, N. (2019). Focusing on Teacher Education to Introduce AI in Schools: Perspectives and Illustrative Findings. *2019 IEEE Tenth International Conference on Technology for Education (T4E)*, 71–77. https://doi.org/10.1109/T4E.2019.00021

Victor, D. G. (2009). *The Politics of Fossil-Fuel Subsidies* (SSRN Scholarly Paper No. ID 1520984). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.1520984

Völcker, C., Molina, A., Neumann, J., Westermann, D., & Kersting, K. (2020). DeepNotebooks: Deep probabilistic models construct python notebooks for reporting datasets. *Communications in Computer and Information Science*, *1167 CCIS*, 28–43. https://doi.org/10.1007/978-3-030-43823-4_3

Wang, D., Ram, P., Weidele, D. K. I., Liu, S., Muller, M., Weisz, J. D., Valente, A., Chaudhary, A., Torres, D., Samulowitz, H., & Amini, L. (2020). AutoAI: Automating the End-to-End AI Lifecycle with Humans-in-the-Loop. *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, 77–78. https://doi.org/10.1145/3379336.3381474

Wang, K., Guo, P., Xin, X., & Ye, Z. (2017). Autoencoder, low rank approximation and pseudoinverse learning algorithm. *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 948–953. https://doi.org/10.1109/SMC.2017.8122732

Ward, S. J. (2009). Journalism ethics (K. Wahl-Jorgensen & T. Hanitzsch, Eds.). *The handbook of journalism studies*, 295–309.

Weidele, D. K. I., Weisz, J. D., Oduor, E., Muller, M., Andres, J., Gray, A., & Wang, D. (2020). AutoAIViz: Opening the blackbox of automated artificial intelligence with conditional parallel coordinates. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 308–312. https://doi.org/10.1145/3377325.3377538

West, J., & Gallagher, S. (2006). Challenges of open innovation: The paradox of firm investment in open-source software. *R&D Management*, *36*(3), 319–331. https://doi.org/10.1111/j.1467-9310.2006.00436.x

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration*, *42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Wischmeyer, T. (2020). Artificial Intelligence and Transparency: Opening the Black Box. In T. Wischmeyer & T. Rademacher (Eds.), *Regulating Artificial Intelligence* (pp. 75–101). Springer International Publishing. https://doi.org/10.1007/978-3-030-32361-5_4

Wong, P.-H. (2020). Democratizing Algorithmic Fairness. *Philosophy & Technology*, *33*(2), 225–244. https://doi.org/10.1007/s13347-019-00355-w

Wu, N., & Silva, E. A. (2010). Artificial Intelligence Solutions for Urban Land Dynamics: A Review. *Journal of Planning Literature*, *24*(3), 246–265. https://doi.org/10.1177/0885412210361571

Xanthopoulos, I., Tsamardinos, I., Christophides, V., Simon, E., & Salinger, A. (2020). *Putting the Human Back in the AutoML Loop*.

Yigitcanlar, T., & Cugurullo, F. (2020). The Sustainability of Artificial Intelligence: An Urbanistic Viewpoint from the Lens of Smart and Sustainable Cities. *Sustainability*, *12*(20), 8548. https://doi.org/10.3390/su12208548

Ylipulli, J., & Luusua, A. (2019). Without libraries what have we? Public libraries as nodes for technological empowerment in the era of smart cities, AI and big data. *Proceedings of the 9th International Conference on Communities & Technologies - Transforming Communities*, 92–101. https://doi.org/10.1145/3328320.3328387

Zuboff, P. S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books.

Appendix A Source documents

Table 4: Literature review documents

Title	Author	Year
A paradigm for democratizing artificial intelligence research	Moreau et al.	2019
AutoAI: Automating the end-to-end AI lifecycle with humans-in-the-loop	D. Wang et al.	2020
Democratizing cognitive technology: a proactive approach	Ienca	2019
The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms	Sudmann	2020
Towards a global participatory platform	Buckingham Shum et al.	2012
AI, Democracy and the Law	Djeffal	2020
Democratization of AI Using Cognitive Services	Masood and Hashmi	2019
Democratizing AI	Allen et al.	2019
Democratization and Impacts of AI	Luce	2019
How will "democratization of artificial intelligence" change the future of radiologists?	Kobayashi et al.	2019
On the media-political dimension of artificial intelligence: Deep learning as a black box and OpenAI	Sudmann	2018
PANDA: facilitating usable AI development	Gao et al.	2018
The sustainability of artificial intelligence: An urbanistic viewpoint from the lens of smart and sustainable cities	Yigitcanlar and Cugurullo	2020
Artificial intelligence and human development: toward a research agenda	Smith and Neupane	2018
Distributed, decentralized, and democratized artificial intelligence	Montes and Goertzel	2019
Artificial Intelligence and Community Well-being: A Proposal for an Emerging Area of Research	Musikanski et al.	2020
Considerations for a more ethical approach to data in AI: on data representation and infrastructure	Baird and Schuller	2020
Better Future through AI: Avoiding Pitfalls and Guiding AI Towards its Full Potential	Miikkulainen et al.	2019
The open-closed principle of modern machine learning frameworks	Braiek et al.	2018
Human rights and technology: New challenges for justice and accountability	Land and Aronson	2020
Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role?	Carter	2020
Towards the collaborative development of machine learning techniques in planning support systems – a Sydney example	Lock et al.	2020
How to Democratise and Protect AI: Fair and Differentially Private Decentralised Deep Learning	Lyu et al.	2020
Constitutional democracy and technology in the age of artificial intelligence	Nemitz	2018
Artificial intelligence and policy: quo vadis?	Lauterbach	2019
Artificial intelligence and the public sector—applications and challenges	Wirtz et al.	2019
An AI race for strategic advantage: rhetoric and risks	Cave and OhÉigeartaigh	2018
Responsible AI: requirements and challenges	Ghallab	2019
The Democratization of Artificial Intelligence: One Library's Approach	Finley	2019
Design and Analysis of Experiments: A Challenge Approach in Teaching	Pavao et al.	2019
Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence	Robinson	2020
Focusing on Teacher Education to Introduce AI in Schools: Perspectives and Illustrative Findings	Vazhayil et al.	2019
Towards an Inclusive Future in AI: A Global Participatory Process	Belinchon et al.	2019
Learning machine learning with very young children: Who is teaching whom?	Vartiainen et al.	2020
A framework for social fairness; Insights from two algorithmic decision-making controversies in The Netherlands	Timan and Grommé	2020
Without libraries what have we?: Public libraries as nodes for technological empowerment in the era of smart cities, AI and big data	Ylipulli and Luusua	2019
AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations	Floridi et al.	2018
Democratizing data science through data science training	Van Horn et al.	2017
A framework for evaluating barriers to the democratization of artificial intelligence	Garvey	2018
The future of human creative knowledge work within the digital economy	Holford	2019
Hard choices in artificial intelligence: Addressing normative uncertainty through sociotechnical commitments	Dobbe et al.	2020
Towards a human rights-based approach to AI: Case study of apprise	Thinyane and Sassetti	2020
Democratizing AI: Non-expert design of prediction tasks	Bagrow	2020
The Democratization of Machine Learning Features	Patel	2020
Algorithmic accountability	Shah	2018
Characteristics of collaboration in the emerging practice of open data analysis	Choi and Tausczik	2017
The Invisible Power of Fairness. How Machine Learning Shapes Democracy	Beretta et al.	2019
Beneficial artificial intelligence coordination by means of a value sensitive design approach	Umbrello	2019
On the promotion of safe and socially beneficial artificial intelligence	Baum	2017
Principles and business processes for responsible AI	Clarke	2019
Democratic governance of digital platforms and artificial intelligence? Exploring governance models of china, the us, the eu and mexico	Schneider	2020
Big tech and societal sustainability: an ethical framework	Arogyaswamy	2020
Artificial Intelligence and Human Rights	Donahoe and Metzger	2019
Artificial Intelligence and Human Rights: A Business Ethical Assessment	Kriebitz and Lütge	2020
Artificial Intelligence and Transparency: Opening the Black Box	Wischmeyer	2020
DLHub: Model and data serving for science	Chard et al.	2019
Machine learning interpretability: A survey on methods and metrics	Carvalho et al.	2019
Towards an integrated solution for intelligent visual	Lee	2020
Towards interactive curation & automatic tuning of ML pipelines	Binnig et al.	2018
Northstar: An interactive data science system	Kraska	2018
Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers	Hohman et al.	2019
A system for accessible artificial intelligence	Olson et al.	2018
kGCN: a graph-based deep learning framework for chemical structures	Kojima et al.	2020
Autoencoder, low rank approximation and pseudoinverse learning algorithm	K. Wang et al.	2017
AutoAIViz: Opening the blackbox of automated artificial intelligence with conditional parallel coordinates	Weidele et al.	2020
Putting the human back in the AutoML loop	Xanthopoulos et al.	2020
Evolutionary neural automl for deep learning	Liang et al.	2019
Democratizing Data Science through Interactive Curation of ML Pipelines	Shang et al.	2019
Slice finder: Automated data slicing for model validation	Chung et al.	2019
DeepNotebooks: Deep probabilistic models construct python notebooks for reporting datasets	Völcker et al.	2020

Appendix B CoCoAI4Privacy

To clarify how the CoCoAI framework can be implemented, I will present an imaginary project called CoCoAI4Privacy, which uses machine learning to predict privacy labels based on privacy agreements. The aim of the project is to develop an AI model able to extract information from privacy agreements. More specifically, it will extract information about how user data is being collected and processed by a company, according to their privacy policy. The goal is to present this information to the user, informing them about the terms they are accepting when registering for the service. The project leverages CoCoAI to address several of the challenges of democratising AI, most notably the access to data, transparency and reducing the barrier of entry to use existing AI.

The CoCoAI4Privacy project will be openly available on GitHub, both the project platform and the resulting AI algorithm, thereby using GitHub as it's **Open-source AI code repository**. This openness enables others to reuse or expand the platform for their own projects. The project platform also includes **Crowdsourced data labelling** and an interface for downloading the dataset (as a partial implementation of the **Open-source dataset repository**). These components enable community members to help increase the size of the dataset by labelling data, and to use the dataset for their own projects. A restricted version of **Distributed computing** is used to train models as the dataset is expanded or changes are submitted to the AI algorithm, keeping track of the versions of the dataset and the algorithm each model was trained with. However, the distributed computing will be limited to authorised contributors, to avoid running code from unknown sources. Finally, there will be an interface where people can submit URLs to privacy policies to have them analysed, which is enabled by a online **Open AI model API** that can be used by others in their own projects. Perhaps, for example, one community member wants to make a smartphone app for scanning privacy policies in the app store.

In summary, one can view CoCoAI4Privacy as a partial implementation of the CoCoAI framework, using the Open-source AI code repository, Crowdsourced data labelling, Open AI model API and a partial implementation of the Distributed computing and Open-source dataset repository.

B.1 Architecture goals

The architecture of this platform should be designed with some special considerations in mind. First and foremost, in order for the platform components to be reusable, the platform has to be flexible. Not all projects will be using the same components, some projects may introduce new components, some may choose not to use all the components and others may replace some components with custom implementations.

Projects should also be able to process different types of data, from images and text, to videos and graphs. The goal is not to provide an implementation for all possible data types, but rather designing each component in such a way that implementing support for a new data type will require touching as few components as possible. In other words, the platform needs

to be easy to change, so it can be adapted and customised as needed.

Furthermore, the architecture should be easy to scale. Supporting small projects hosting several of the components from the same server, to dynamic large-scale projects that spin up new instances of components in order to scale horizontally to meet demand.

This dynamic nature also requires that the architecture enables debugging, as well as issue and performance tracking across separate components. So any problems can be traced, following the request as it is processed by each component. Further, it should support status, health and performance tracking for the different components to show whether any component is throwing errors or running near max capacity.

Finally, the dynamic nature of this platform demands support for two different types of configuration of the services. First, the platform should enable configuration through simple configuration files, for smaller projects using only one or a few of the components. Second, the platform should support centralised configuration where each component retrieves its configuration from a configuration service, a single source of truth. This enables project administrators to change the configuration in one place and every related service instance will automatically update to use the new settings.