Artificial Intelligence, Ethics and Society

An Overview and Discussion Through the Specialised Literature and Expert Opinions

OBSERVATORI D'ÈTICA EN INTEL·LIGÈNCIA ARTIFICIAL DE CATALUNYA

Generalitat de Catalunya
Departament de la Vicepresidència
i de Polítiques Digitals i Territori

i2cat

Universitat
de Girona

Artificial Intelligence, Ethics and Society: An Overview and Discussion Through the Specialised Literature and Expert Opinions

September 2021

Edition

Observatori d'Ètica en Intel·ligència Artificial de Catalunya (OEIAC)

Carrer de la Universitat, 10 - Edifici Econòmiques

Universitat de Girona 17003 Girona

Authors

Albert Sabater (OEIAC) and Alicia de Manuel (OEIAC).

Acknowledgements

Daniel Santanach (Direcció General d'Innovació i Economia Digital, Generalitat de Catalunya) and Àngel Martín Carballo (Fundació i2CAT).

Artificial Intelligence, Ethics and Society

An Overview and Discussion through the Specialised Literature and Expert Opinions

2021

OBSERVATORI D'ÈTICA EN INTEL·LIGÈNCIA ARTIFICIAL DE CATALUNYA

Index

Foreword

In recent years we have seen how artificial intelligence has become an increasingly ubiquitous technology in our daily lives. Systems for voice recognition, problem solving, learning and planning or decision support are some examples of the digital transformation we are undergoing. The most important advances in artificial intelligence research in recent years have been in the field of machine learning. Specifically in the field of deep learning, and this has been driven in part by greater data availability and an exponential increase in computing power.

However, we could say that the technology of artificial intelligence systems is still at an early stage, technically known as “narrow AI”, due to the limitations of systems to adapt and improvise in new environments and to apply it in unfamiliar scenarios. But it seems obvious that it is only a matter of time before a large part of artificial intelligence systems will enable us to increase our capabilities and help us even more as a society. And this undoubtedly represents a great opportunity, as demonstrated by the fact that in Catalonia, Europe and the rest of the world are devoting and investing a great deal of efforts and resources to the development of artificial intelligence.

But we know that the widespread use and implementation of artificial intelligence also entails numerous ethical challenges such as responsibility, justice, transparency and privacy, which must be taken into account so that its development generates trust in society as a whole. It is for this reason that the Artificial Intelligence Strategy promoted by the Generalitat de Catalunya, under the name of CATALONIA.AI, has initiated a programme of actions to strengthen the Catalan ecosystem in artificial intelligence, with a focus on “Ethics and Society” to promote the development of ethical artificial intelligence, which respects the law, is compatible with our social and cultural norms, and is people-centred.

Within this context, thanks to the collaboration between the

Department de la Vicepresidència i Polítiques Digitals i Territori of the Generalitat de Catalunya and the Universitat de Girona, the Observatori d'Ètica en Intel·ligència Artificial de Catalunya, aka OEIAC, was born on 29 June 2020, taking the form of a chair at the Universitat de Girona. The main objective of OEIAC is to study the ethical, social and legal consequences and the risks and opportunities of the implementation of artificial intelligence in everyday life in Catalonia from a fully transversal perspective and with an operational structure that acts as a bridge between humanism, science and technology and that guarantees the dynamics of the quadruple helix.

This report is a good example of how the OEIAC positions itself with the will to coordinate and structure a reflection on the ethical considerations and social impact of artificial intelligence. For this reason, I would like to thank all those who have contributed with their knowledge and extensive analysis of the specialised literature.

Jordi Puigneró i Ferrer
Vicepresident del Govern i conseller de Polítiques Digitals i Territori

Introduction

Safeguarding our autonomy as humans in decision-making in the midst of large-scale digitalisation, and the development and implementation of artificial intelligence (AI) in almost all activities of our daily lives, may seem out of sync with reality. However, at the Observatori d'Ètica en Intel·ligència Artificial de Catalunya we believe that this is not only possible but essential to ensure that these new tools, the so-called AI systems, are people-centred and their use is characterised by prioritising principles such as transparency, justice, accountability and privacy, among others. The increasing use of AI systems in different spheres, from professional to personal, and in our social interactions, raises questions about some of the main social covenants on which community life is based, and raises several social and economic issues, for example, in relation to the future of work or the distribution of wealth. In this sense, the priority for any public institution or company should be to clearly identify what these challenges are in order to turn technological innovation into a shared and well-defined vision on the ethical and social shaping of AI in practice.

This is the idea behind this paper, which aims to provide an overview of the ethical and societal implications of AI through a review of specialised literature and expert opinions. Taking into account the developments and the main issues raised in academia, industry and policy, the report aims to provide useful information and elements for the debate around the opportunities and limitations of AI.

For this reason, in order to carry out the first part of this work, we have conducted

OEIAC
Observatori d'Ètica en Intel·ligència Artificial de Catalunya

The Observatory, which takes the form of a Chair at the Universitat de Girona, has as its main objective to study the ethical, social and legal consequences and the risks and opportunities of the implementation of Artificial Intelligence in everyday life in Catalonia. Its creation in 2020 is part of the Artificial Intelligence (AI) Strategy promoted by the Government of Catalonia under the name CATALONIA.AI.

an extensive review of the specialised literature, covering more than 130 publications including academic and scientific texts, technical or research reports from government agencies, companies and associations, as well as newspaper articles. In the second part, we have addressed the development and impact of AI from different fields, from academia to industry, public administration and citizenship, through a total of 23 interviews that allow us to know various opinions and reflections on the current development and implementation of AI systems taking into account different ethical, social and legal aspects.

While the aim of this report is not to set out a roadmap on the ethical and social development of AI, it does identify key research directions that prioritise building a shared knowledge base and discourse that can underpin an ethical and social approach. AI ethics is an emerging field that seeks to address the new risks posed by AI systems. While the field is currently dominated by a proliferation of AI 'codes of ethics' that seek to guide the design and deployment of AI systems, these are limited in key respects as they lack a universally agreed framework and are not binding as specific legislation. But sometimes it goes much further, as those who design and implement AI tools do not always put them into operation and, too often, do not take into consideration the people potentially affected by AI systems. Indeed, this governance model relies heavily on corporate self-regulation, a worrying prospect given the absence of democratic representation and social accountability in corporate decision-making.

There is an urgency at the moment, and that is to work towards making ethical and social considerations central to the use and implementation of algorithms and data in various AI systems. As we discuss how to apply ethical principles and human rights to AI systems, digital technologies continue to evolve rapidly. In this context, the COVID-19 pandemic is a timely reminder that to ensure that AI tools advance people's progress and well-being, it is critical to be proactive and inclusive in the development of such tools. We know that AI systems can be extremely powerful, generating analytical and predictive insights that progressively surpass human capabilities. This means that they are likely to be increasingly used as substitutes for human decision making, especially when the analysis needs to be done very quickly with the aim of gaining efficiency.
But we should not forget that, also in these terms, we can cause serious harm to individuals or groups who are vulnerable, especially when AI systems do not provide for representativeness and are considered infallible. In this sense, this report not only deals with descriptive issues around the main ethical principles in AI, but explores these emerging ethical considerations, exposing some of the main societal issues and perceptions, as well as the main institutional and business responses.

Clarifying the multiple issues and resolving tensions between principles, values and actors is crucial if AI systems are to be developed and used for the benefit of society. It should be noted that the review and analysis in this paper is not exhaustive and is necessarily based on a limited and shared societal understanding of both the technological issues surrounding AI and the ethical, social and legal aspects that can underpin a humanistic approach to AI. While there is a growing consensus on fundamental issues such as responsibility, social justice, privacy and consent, we know that these can be subject to different meanings in different contexts, just as ethical values and social impact are subject to different societal definitions, norms and values. In this sense, the paper argues that the ethics of AI is as much about the process itself as it is about the outcome, which requires a scientific understanding of the world and working closely with affected people.

Taking into account the need to move towards a more ethical and social AI, in this paper we wanted to capture the more present vision, but we also addressed a part of the prospective or future vision, mainly through the qualitative information obtained from the interviewees. Possibly the vision of the future in its myriad forms, fascinating or alarming, bright or gloomy, says more about our fantasies and fears than about the future itself. Nevertheless, visions of the future generate a power to look ahead, and also allow us to appeal to all citizens to become users and critics of technologies in general and of AI in particular. At the Observatori d'Ètica en Intel·ligència Artificial de Catalunya, we believe that this is a fundamental aspect to identify and resolve tensions between different interest groups, and to build a more rigorous evidence base to promote the opportunities of AI while debating the ethical and social issues surrounding the use and implementation of AI.

PART 1

1.1. What do we mean by artificial intelligence (AI)?

"THE HUMAN SPIRIT MUST PREVAIL OVER TECHNOLOGY"
ALBERT EINSTEIN

In recent decades, the technology sector has experienced tremendous growth and, in particular, AI systems are benefiting from what is known as a summer period, namely a time when large public and private funding brings to the surface new opportunities from these systems, but also new challenges and problems in their application.

The technological and highly technical nature of AI has meant that the technology has traditionally been confined to expert and specialist circles, so it is not surprising that, nowadays, the mere mention of AI is enough to imply that its use involves a digital innovation that is systematically associated with institutions or companies wishing to project an attractive and futuristic image. But beyond the technical realities and projects it is supposed to denote, the main problem we have today is the need for a more precise definition to clarify the public debate in order to avoid the representation of a new mythology of our time.

So what do we mean by AI? Broadly speaking, AI can be defined as "the science and engineering of making intelligent machines, especially intelligent software. It is related to the similar task of using computers to understand human intelligence, but AI need not be limited to methods that are biologically observable" (McCarthy, 2007: 2). In that sense, is AI trying to simulate human intelligence? John McCarthy himself specifies that "sometimes, but not always or even usually. On the one hand, we can learn something about how to make machines solve problems by observing other people or simply by observing our own methods. On the other hand, most of the work in AI involves studying the problems that the world presents to intelligence rather than studying people or animals. AI researchers are free to use methods that are not observed in people or that involve much more computing than people can do" (McCarthy, 2007: 3). John McCarthy first coined the term AI in 1956, when he invited a group of researchers from various disciplines, such as language simulation,

neural networks and complexity theory, to a summer workshop called the Dartmouth Summer Research Project on AI, to discuss what would eventually become the field of AI. However, before that, there is the precedent of Alan Turing and his 1936 paper *On Computable Numbers, with an Application to the Entscheidungsproblem* and his 1950 paper *Computing Machinery and Intelligence*, which would lead to the further development of all computer science, especially from his formulation and question: "Can machines think?" From there, Alan Turing himself developed what is known as the "Turing Test", in which a human interrogator would try to distinguish between a computer and a human text response. Although this test has been the subject of much scrutiny since its publication, it remains an important part of the history of AI. Later and from a more contemporary point of view, Stuart Russell and Peter Norvig (2003) edited the book *Artificial Intelligence: A Modern Approach*, which highlights four possible goals or definitions of AI systems on the basis of rationality and thinking versus acting: (1) systems that think like humans, (2) systems that act like humans, (3) systems that think rationally, and (4) systems that act rationally.

A more encyclopaedic definition is provided by Copeland (2021) who describes AI as "the ability of a computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is often applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalise, or learn from past experience". Today, modern dictionary definitions tend to focus on how AI can mimic human intelligence. For example, the *Oxford Living Dictionary or Lexico* gives this definition: "the theory and development of computer systems capable of performing tasks that normally require human intelligence, such as visual perception, speech recognition, decision making and translation between languages"[1] . On the other hand, and with a focus on present capabilities, the *European Commission's White Paper* defines AI[2] , following the recommendations of the High-Level Expert Group on AI, as follows:

"Artificial intelligence (AI) systems are human-designed software (and possibly also hardware) systems that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the structured or unstructured data collected, reasoning about the knowledge, or information processing, derived from this data and deciding the best action(s) to take to achieve the given goal. Artificial intelligence systems can use symbolic rules or learn a numerical model, and can also adapt their behaviour by analysing how the environment is affected by their previous actions'"(European Commission, 2020: 16).

[1] https://www.lexico.com/definition/artificial_intelligence (accessed 27 August 2021).

[2] European Commission (2020). White Paper on Artificial Intelligence: A European approach to excellence and trust (COM(2020) 65). Brussels: European Commission, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (accessed 27 August 2021).

AI is certainly not a monolithic term as some of the definitions of AI we have collected demonstrate. Moreover, AI requires nuance whether we analyse it through its evolutionary stages or focus on different types of systems, such as analytical AI, human-inspired AI or humanised AI (Kaplan and Haenlein, 2019). We know that the progress of the history of AI from the 1950s to the present day has not been continuous, and although the ultimate goal of many computer scientists and engineers has been to build AI systems that are indistinguishable from human intelligence, many people researching in this field have been forced to divert their initial focus from developing machines that behave in a way that would be considered intelligent to a human towards more specific tasks. For this reason, when we talk about AI we refer among other things to solving problems such as image recognition, natural language understanding or games (e.g. chess or checkers). However, we should bear in mind that any definition of AI may seem questionable depending on the point of view. By using a definition close to that of the European Commission, we seek a minimum and operational basis for discussion, which is a prerequisite for pragmatically outlining the scope of AI systems, especially in relation to ethical and social impact issues. In other words, it is a matter of using a definition that, while not the most precise possible, takes into account the reasons why we care about AI today. However, as we will see later, the importance of a definition of AI also lies in the fact that in any new legal instrument, the definition of AI will have to be flexible enough to adapt to technical progress and, at the same time, be precise enough to provide the necessary legal certainty. Not surprisingly, the definition of AI has changed over time.

For this reason, it is also important to take into consideration the definitions of AI in terms of the objectives that an AI system is trying to achieve, which, in general, and taking into account the work of Russell and Norvig (2003), can be grouped in the following three groups: (1) to build systems that think exactly like humans; (2) to get systems to work without figuring out how human reasoning works; and (3) to use human reasoning or behaviour as a model, but not necessarily as the end goal. Arguably, most of the AI development being undertaken by industry leaders today falls into the third group as it uses human behaviour as a guide to deliver better services or create better products, rather than trying to achieve a perfect replica of the human mind. In the quest for achieving this, there has been a transition from the symbolic approach to the connectionist approach. In the former, also known as symbolic AI, procedural computational algorithms are combined with Boolean mathematics and logic systems to obtain knowledge representations that can then be used as reasoning algorithms. However, as this type of symbolic AI suffers from a certain kind of rigidity, as programmers manually elaborate basic rules and the syntax itself, a family of algorithms known as connectionist AI has been developed. This approach, which tries to mimic the neural

structure of the brain (Deep Neural Networks), aims to perform better than symbolic AI by generalising well and learning from examples. However, the solutions provided often lack interpretability, as the rules and operation of this type of AI using differential equations do not have an inherent connection to the learned representation, but work together as a whole. This means that we cannot always understand why they work in a certain way, a fact that has meant that connectionist AI is also often referred to as a black box.

Indeed, so-called connectionist AI based on the workings of the human brain and its interconnected neurons has received much more attention in recent years, thanks in part to the advent of big data. Connectionist AI is generally considered to be a good choice when a lot of high-quality training data is available to feed the algorithm. Although connectionist AI models become "smarter" with increased exposure, an accurate information base is always needed to initiate the learning process. Although both approaches have been used since the birth of AI, the symbolic approach was dominant for much of the 20th century, but today the connectionist approach is clearly ascendant, mainly machine learning with deep neural networks. Both paradigms are considered to have strengths and weaknesses, and the most important challenge for the current AI field is to reconcile the two approaches (Garnelo and Shanahan, 2019).

Taking both approaches into account, we are currently at an early stage of AI that is commonly referred to as narrow AI or artificial intelligence, a term that describes current AI systems as only being capable of performing a specialised task. The next evolutionary step for these systems would be called artificial general intelligence (AGI) where the AI would be able to emulate human intelligence and would therefore be able to perform any intellectual task that a human can perform. An artificial superintelligence (ASI) would be an AI that would surpass human capabilities and therefore could give way to what is called the singularity, which is this idea whereby the trajectory of AI would reach a point where they themselves could develop more AI systems that surpass the level of human intelligence. In this approach, some interesting debates arise whereby the idea is that AI systems will be far from human control and difficult to predict. This could then give rise to the problem of value alignment whereby we ask to what extent we can be sure that a super-intelligent AI system will have a positive resolution in accordance with human perception.

Although we are still at an early stage of AI, there have been a number of milestones in recent years, such as the victory of Alpha Go (Google) over world Go champion Lee Sedol in March 2016. While this victory is usually presented as an achievement in symbolic terms, it is worth noting that, unlike chess, Go does not lend itself to memorising a large

number of moves that the machine could simply reproduce, but to a large number of possible combinations. In fact, Alpha Go's victory illustrates the fact that recent advances in AI are due, above all, to the development of Machine Learning, which is one of its most notable applications. In fact, before the emergence of Machine Learning, programmers had to divide the task to be automated into multiple instructions, so that all steps had to be clearly specified in order to perform a task. In machine learning, the main difference is that there are not multiple instructions but a system that makes its own decisions based on a lot of data from which it can learn and make its own decisions. In this sense, this technique allows much more complex tasks to be performed than a conventional algorithm because systems can act without being explicitly programmed. Machine learning-based AI therefore refers to algorithms that have been specifically designed so that their behaviour can evolve over time, depending on the input data.

Deep learning, on the other hand, is a subclass of deep learning and is the cornerstone of recent advances in machine learning. Within deep learning we can distinguish between supervised learning (when the input data used by the system is labelled by humans) and unsupervised learning (when the input data is not labelled by humans and it is an algorithm that performs its own classification). In supervised learning, people are required to teach the machine the result to produce, i.e. they must "train" it. To do this we can supply the machine with information such as thousands of photographs that have previously been labelled with, for example, the elephant identifier, along with others that have been explicitly labelled with an identifier that they are not elephants. Platforms such as Amazon's Mechanical Turk, which recruits thousands of workers to classify thousands of photographs that are then used to train an image recognition programme, or Google's reCAPTCHA system are examples of more ambitious supervised learning systems. Beyond shape recognition in images, learning systems can also classify other types of information. If the goals are real-life applications, such as autonomous driving, action recognition, object detection and recognition in live broadcasts, then systems need to be trained on video data.

And of course, in addition to images and video, we can also use text information such as spam among incoming email messages. In fact, Gmail is a simple and typical example of AI in practice as Google gathers a considerable and constantly updated database of spam reported by its users. The system uses information to learn to identify what characterises a spam message and can thus decide for itself which messages to filter and classify between spam and non-spam. On the other hand, the goal of unsupervised learning systems (much less common) is to train a network of models (called "discriminators" or "encoders") for use in other tasks. The characteristics of these models need to be general enough to be used in categorisation tasks, so that they can

provide equivalent results to those obtained by supervised systems. However, to date, supervised models always perform better than unsupervised pre-trained models. This is because supervision allows the model to better encode the characteristics of the data set. But supervision can also be decreasing if the model is then applied to other tasks. In this sense, it is expected that unsupervised training can provide more general features for learning to perform any task.

As AI systems become more widely deployed in our daily lives and its aplication in private and public spheres is greater, the ethical considerations and the social impact of the use of AI have also been growing. Indeed, in addition to the ethical concerns of its use and the consequences it may have on people and their environment, there are also other concerns about the use and exploitation of the data that feed AI systems.

1.2. What do we mean by the ethics of AI?

While definitions of AI focus on a technological description of the use of machines such as computers or robots to do things that would normally require human intelligence, it is important to underline that AI can also be defined by its implications, especially in how it has affected society and the ways in which we interact with each other. In that sense, while AI can do a lot of good (for example, by making products and processes safer and more efficient), it can also cause a lot of harm. Harm that can be both material (e.g. in terms of people's safety and health, including loss of life, and damage to property) and immaterial (e.g. loss of privacy, limitation of the right to freedom of expression and human dignity, or discrimination).

For this reason, when we talk about the ethics of AI, we mainly refer to two concerns. The first concerns the moral behaviour of people in the design, manufacture and use of AI systems (ethics of technology), and the second concerns the behaviour of AI systems (machine ethics). Regarding the first concern, it is worth noting that the mere fact that a machine is designed by one or more people and attempts to "mimic" human intelligence can already raise a number of problems such as deception, biases and cognitive errors, which are likely to appear systematically if the AI is built to resemble people (Boden et al., 2017; Nyholm and Frank, 2019).

For this reason, the application of ethics in AI is crucial, among other things, to ensure that it does not harm people and their environment, as well as other living beings and their habitats. It is worth stressing at this point that, within the concern of AI ethics, the central question is not whether AI systems can do one thing or another, the question is whether and how they should do it. Therefore, AI ethics is concerned, among other things, with solving or mitigating problems related to bias, providing safety guidelines that can prevent undesirable risks (even of an existential nature for humanity) and, ultimately, building AI systems that by adopting ethical norms can help us move forward as a society.

As for the second concern, it would be closely linked to the idea that "the greater the freedom of a machine, the more it needs moral rules" (Picard, 1997: 19), which would mean that all interactions between AI systems and people necessarily involve an ethical dimension. In some ways, the idea of implementing ethics within AI systems is one of

the main research goals in the field of AI ethics (Lin et al., 2012; Wallach and Allen, 2009). It is argued that responsibility has increasingly shifted from humans to autonomous AI systems, and that these are able to work much faster and more efficiently than humans (without taking breaks and without the need for constant supervision).

Generally, we can say that within the ethics of technology we also find the sub-disciplines of robot ethics (Gunkel, 2018; Nyholm, 2020), which deals with issues related to how humans design, build and use robots; and computer ethics which focuses on aspects of information processing through computers (Johnson and Nissenbaum 1995; Himma and Tavani 2008), and where aspects such as data security or privacy issues prevail. Therefore, we can say that AI ethics seeks to resolve questions of human morality within AI systems, taking into account that they can be used to do good and evil, and taking into consideration moral principles such as responsibility, justice, trust, transparency, inclusiveness and sustainability, among others.

Numerous approaches to implementing ethics within AI systems have been proposed over the past two decades in order to provide these systems with principles that they can use to make moral decisions (Gordon, 2020a). In this regard we can distinguish at least three types of approaches: a top-down approach (theory-based reasoning), a bottom-up approach (shaped by evolution and learning), and a hybrid approach, in which both approaches are considered as the basis for moral reasoning and decision-making. The first approach is based on an initial learning process of the AI system through specific programming that would allow the system the ability to solve (new) ethical dilemmas on its own. An example would be Guarini's (2006) system, which bases its ethical decisions on a learning process in which a neural network is presented with known correct answers to ethical dilemmas. After the initial learning process, the system is supposed to be able to solve new ethical dilemmas on its own. However, Guarini's system generates problems with the reclassification of cases, caused by the lack of adequate reflection and accurate representation of the situation. In fact, Guarini admits that casuistry alone is insufficient to implement ethics in AI systems.

The second approach combines two main ethical theories, utilitarianism and deontology. Utilitarian reasoning is applied until principles or values are affected, at which point the system operates in deontological mode and becomes less sensitive to the utility of actions and consequences. To align the system with human moral decisions, authors such as Dehghani et al. (2011) propose to evaluate it on the basis of psychological studies on how most human beings decide specific cases. Although this approach is considered particularly sound because it adequately respects the two main ethical theories (deontology and utilitarianism), their additional strategy of

using empirical studies to reflect human moral decisions may be problematic. In this sense it is criticised because it may consider as correct only those decisions that align with the majority opinion, which may be misleading and as such may have serious consequences. For this reason this approach is seen as a descriptive model for the study of ethical behaviour, but not as a model of normative ethics.

The hybrid approach combines a top-down component (theory-based reasoning) as well as a bottom-up component (shaped by evolution and learning). The model presented by Wallach et al. (2010) is not necessarily inaccurate with respect to how moral decision-making works in an empirical sense, but their approach is descriptive rather than normative in nature. Therefore, their empirical model does not solve the normative problem of how AI systems should act. It should be stressed that descriptive ethics and normative ethics are two different things, since the former tells us how humans make moral decisions while the latter is concerned with how we should act.

But the ethics of AI is not just a question of design and moral implementation in systems, it is also a question to be assessed in terms of social and cultural values, which one would expect to be embedded in AI designs. Therefore, in considering how development and design affect the ethics of AI across cultures, we must ask: who does or participates in AI design, and for whom is it done?

But the persistent global digital divide prevents people in many parts of the world from participating in the design and development of AI technologies. For example, in many parts of the world, people lack the educational opportunities needed to acquire the skills to develop AI systems, meaning that population groups (especially women, ethnic minorities and other vulnerable groups) are particularly affected by this skills digital gap. This implies that the logic of AI design is very much bounded not only by ethical issues but also by an intercultural perspective that is deeply intertwined with society and its inequalities. From this perpective and from an ethical point of view, what are the long-term social consequences of AI systems being developed without the full participation of women, ethnic minorities and vulnerable groups?

There is a close relationship between the social and the technological that determines how AI is designed but, more importantly, how it is used. Technology is rarely used in laboratory conditions or by people with the same demographic profile as those who designed or tested it. It is important to note that AI systems are used in different societies built over centuries of history and with particular economic and political structures. For this reason, when we talk about the ethics of AI we must also take into consideration that the most carefully designed technologies can function imperfectly

and problematically in the real world, generating an impact that is not always desired.

As we will see below, the approach to these and other concerns in the form of guidelines or regulations is carried out from different institutional frameworks. However, the lack of a global consensus is posing one of the greatest challenges to the adoption of such measures, partly because perceptions, attitudes, discussions on the acceptance and use of AI vary across regions and countries of the world. Different cultural norms and values prevailing in different societies around the world as well as economic models and legislative, executive and judicial characteristics, not only shape the state of AI technology, but also have an impact on the degree of adoption of ethical AI.

As a consequence, we see how the adoption of ethical AI is still irregular and polarised along three main axes that coincide with the main powers of technological development. Thus we speak of an "AI for control" model characterised by the use of AI systems as tools for social and security control. This model is dominated by China's technological developments, through, for example, the implementation of the social credit system or its China 2025 technological development plan. It should be noted that the Asian powerhouse ranks second in the world in the development of AI, and approximately 11% of all AI companies are located in China. Kai-Fu Lee (2018) in his latest book *AI Superpowers: China, Silicon Valley and the New World Order*, rightly emphasises how China pursues a policy that prioritises the idea of the greatest good for the greatest number rather than a moral imperative to protec individual rights, which predominates in the West (Ortega, 2020).

Next, we find the "AI for profit" model oriented towards the development and implementation of AI systems where a few companies dominate the majority of the technology sector, as well as a prioritisation of the economic benefit generated by the application of these systems. We can find this predominant model in the United States, which is currently the leader in the development of AI with 40% of the world's AI companies located in its territory, and where we could say that to date ethical considerations appear in the background, giving much more value to economic growth. Furthermore, it is important to note that, in this context, only a handful of companies lead most of the technological development and economic growth, defining the course of the market and further technological advances, which has provoked a series of debates in recent years about control in private entities. It should be noted that, in terms of data regulation, there is no single data protection law, but rather its regulation is governed by hundreds of state and federal laws that attempt to address the multiple aspects.

Finally, the "AI for society" model is the framework that aims to oppose and distance itself from the Chinese and US models by putting user privacy and ethical principles before the technological development of AI. This is the position that the European Union is taking, and to this end, different initiatives have been taken to coordinate and regulate measures for the development of technology and the creation of legislative frameworks. In this sense we can say that Europe currently has the most advanced legislation in terms of personal data and promotes a policy focused on the right of individuals to decide how their data should be used. For example, while in Europe consent must be obtained explicitly and for a defined purpose, in the United States, consent is tacit, which means that if an individual does not want a company to process his or her personal data, he or she will have to specifically demand that it not do so. However, efforts to adopt ethical principles in AI may also mean that technological innovation may be delayed, causing other less restrictive markets to lead technological development and thus not only increase their lead, but further widen the technological gap between global regions.

The rapid development of AI and its evolution in industry have been two factors that have led to this period being called the Fourth Industrial Revolution. This consideration accentuates the speed of the disruptive changes that are taking place in the technological and economic sector and that are influencing the rest of the sectors in terms of its scope, scale and complexity. The implications of this transition, as well as its potential consequences, demonstrate the imperative need to build and build on a clear state, regional and global regulatory framework that can protect both the rights of individuals and their impact on the planet, as well as facilitate a technological transformation that, far from being traumatic, can exploit its benefits to the full.

1.3. The main ethical principles of AI

We could say that the first ethical principles of AI come from the field of science fiction. Thus, it was Isaac Asimov who presented his Three Laws of Robotics in *Runaround* (Asimov, 1942), and these three were later complemented by a fourth law, called the Zeroth Law of Robotics, in *Robots and Empire* (Asimov, 1986).

These laws, which are considered more utilitarian than moral, respond to the programming of a limited AI. Today, Asimov's proposals have changed to embrace more social issues under the idea of inclusive technology, service to people and the possible rights of "sentient" AI.

Along these lines, we find Mashasiro Mori's (1970) "uncanny valley" theory, which argues that creating robots that are too similar to humans would cause a rejection response.

LAWS OF ROBOTICS

1. **A robot may not injure a human being or, through inaction, allow a human being to come to harm;**
2. **A robot must obey the orders given it by human beings except where such orders would conflict with the First Law;**
3. **A robot must protect its own existence as long as such protection does not conflict with the First or Second Law;**
4. **A robot may not harm humanity, or, by inaction, allow humanity to come to harm.**

We have seen how different countries have tried to establish suggestions and regulations on the issue of ethics and AI. This means that different AI developments, which are being carried out to increase the accessibility of AI, have different approaches and may be affected by cultural, social or religious values. For instance, in Japan the religious culture of Shintoism's approach to beings and things where many objects are attributed soul-like characteristics makes for a more natural approach to the relationship that is established with machines. On the other hand, in the African Ubuntu culture, a person is a person through other people and therefore the link with humans is paramount.

Thus, when implementing an AI system, we must take into account the different ethical and cultural approaches of the recipients of these systems and the communities where they are implemented.

According to the first guidelines on ethical AI published by the High-Level Expert Group on AI (HLEG AI) (European Commission, 2019)[3], trustworthy AI must respect all applicable laws and regulations, must respect ethical principles and values, and must be robust both from a technical perspective and must take into account its social environment. In addition, the HLEG AI guidelines present a set of 7 key requirements that AI systems must meet to be considered trustworthy, which are as follows:

1

Human agency and oversight

AI systems must empower human beings, enabling them to make informed decisions and promoting their fundamental rights. At the same time, appropriate oversight mechanisms need to be ensured.

2

Technical robustness and safety

AI systems must be safe, guarantee a backup plan in case something goes wrong, and be accurate, reliable and reproducible. This is the only way to ensure that unintended damage can also be minimised and prevented.

3

Privacy and data governance

In addition to ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account data quality and integrity and ensuring legitimate access to data.

4

Transparency

Data business models and AI systems must be transparent. Traceability mechanisms can help to achieve this. In addition, AI systems and their decisions should be explained in a way that is tailored to the stakeholders in question. Humans should be aware that they are interacting with an AI system and should be informed of the system's capabilities and limitations.

5

Diversity, non-discrimination and equity

Unfair bias must be avoided in the use of AI systems, as it could have multiple negative implications, from marginalising vulnerable groups to exacerbating prejudice and

[3] European Commission (2019) Ethics guidelines for trustworthy AI. Brussels: European Comission, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accesed 27 August 2021)

discrimination. In fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their lifecycle.

6 **Social and environmental welfare**

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

7 **Accountability**

Mechanisms must be put in place to ensure responsibility and accountability for AI systems and their results. Auditability, which enables the evaluation of algorithms, data and design processes, plays a key role in this, especially in critical applications. In addition, adequate and accessible remediation must be ensured.

In addition to the High-Level Expert Group on AI's ethical guidelines for reliable AI, numerous papers, recommendations and frameworks have been published in recent years on what the ethical principles of AI should be. In *The global landscape of AI ethics guidelines* (2019), Anna Jobin shows us the current lack of consensus among the ethical principles in the AI papers reviewed for the study. This lack of consensus may show that the interpretation of the principles demonstrates the interest of organisations and institutions in highlighting different issue areas. However, some vague and unspecific statements may lead to suspicions that ethics is being used as a whitewash or a way of circumventing regulation rather than an examination of ethical practices, as we saw in the previous section.

To the global inventory maintained by AlgorithmWatch we can find more than 70 proposals[4] very close to the recommendations made in the *Barcelona Declaration* (2017) and the *Montreal Declaration* (2018). Likewise, the recent comprehensive studies by Jobin et al. (2019) and Hagendorff (2020) identify the 11 most common values or ethical principles among the existing proposals:

[4] AlgorithmWatch (2020) AI Ethics Guidelines Global Inventory. https://inventory.algorithmwatch.org/ (accessed 21 Juny 2021)

1

Transparency/ explainability

It is presented as a way to minimise risks and identify and correct rights violations through audits of the systems themselves. This is fundamental to the fulfilment of social objectives such as participation, self-regulation and trust. It is also important for demonstrating and justifying the need to use an IA system.

2

Justice/equity

It refers to people having fair access and treatment to AI systems, to data and thus to the benefits of their implementation. It is mainly expressed in terms of fairness and prevention as well as monitoring or mitigation of bias. While some focus on fairness with respect to diversity, inclusion and equality, others use this concept to appeal or challenge decisions made through AI systems in terms of redress.

3

Security/prevention

It responds to the fact that AI should never cause foreseeable or unintended harm. It also identifies specific security risks such as cyber warfare or hacking, and others such as intrusion discrimination or general privacy violations, including the psychological, emotional or economic impact that may result. The security and harm prevention guidelines focus mainly on technical and governance measures for AI.

4

Responsibility

It refers to acting with integrity in AI actions and decisions. It includes issues such as whether AI should be accountable in a human way, or whether humans should be the only actors who are ultimately responsible for AI systems. Ultimately, to clarify the attribution of responsibility, including legal responsibility.

5

Privacy

The concept of privacy is seen as a value to be upheld and also as a right to be protected. Privacy is often presented in relation to data protection, and is also related to freedom and trust. There are three ways to achieve this according to the literature: (1) technical solutions (e.g. design, data minimisation, access control), (2) research and awareness raising and (3) regulations to adapt to AI specifications.

6

Welfare

It is applied in reference to AI that promotes human well-being, resource creation and socio-economic opportunities. There is some uncertainty around which actors will benefit from AI. For example, the private sector tends to highlight the benefits of AI for customers, but the idea of a benefit for all (Sustainable Development Goals or SDGs) is also shared. The concept is regularly used as a strategy that aims to align AI with human values and rights, with scientific understanding of the world and working closely with affected people.

7

Freedom/ Autonomy

Relating to the freedom to use a preferred platform or technology, or to free technological experimentation through technological empowerment. It also refers to the freedom to withdraw consent. Generally autonomy is also related to user control, mainly to protect privacy. It is often believed that freedom and autonomy are promoted through transparency.

8

Trust

Related to the idea of AI developers and organisations applying trustworthy 'design principles' on the one hand, and stressing the importance of customer trust on the other. The use of this principle has become increasingly important, as it is believed that a culture of trust is fundamental to the achievement of organisational goals. The concept is also used to warn against possible overconfidence in AI and the importance of ensuring that AI meets expectations.

9

Sustainability

Relating to the deployment of this technology to protect the environment, enhance ecosystems and biodiversity, contributing to more just and equal societies and promoting peace. Often used to refer to increasing its energy efficiency and minimising its ecological footprint. It is also used to refer to sustainable socio-economic developments, such as the one referring to the relationship between AI and the future of work.

10

Dignity

It is sometimes used to argue that AI should not diminish or destroy people's dignity, but respect, preserve and, above all, enhance it. However, dignity can be preserved if it

is respected by AI developers, including through new legislation, initiatives, technical guidelines and methodologies, some issued by governments themselves.

11

Solidarity

It refers to the consequences of "radical individualism" resulting from the implementation of AI. It also refers to the implications that AI may have on the labour market and social cohesion, especially for the most vulnerable individuals and groups.

1.4. Why the emergence of ethical AI?

Over the last few years, ethical AI has evolved from a philosophical question to a tangible necessity. The proliferation of smartphones and the AI applications we use on a daily basis, the impact of technology in all sectors (including industry, healthcare, justice, transport, finance and entertainment), the increase in data processing capacity, as well as the threat of an arms race of smart weapons, has sparked the debate to establish the principles that make sense of ethical AI.

As we mentioned earlier, we need to situate this emergence within the progress of the so-called Fourth Industrial Revolution, where, alongside AI, we find other technologies such as Big Data, robotics and the Internet of Things (IoT). This convergence of digital technologies is characterised by the speed, scope and impact it is having on the transformation of our society. Interoperable systems and decentralised decisions are part of the design principles along with the transparency and technical support that these systems offer. We must understand that, like other technological developments before it, AI is still a tool applied in different domains. But its rapid development, as well as its application to a wide range of sectors of society, has led to the emergence of different ethical challenges in the face of the uncontrolled advance of this technology. O'Neil (2016) alerts us to how the advance of predictive models in AI in general have become *Weapons of Math Destruction*, since, although these systems are designed as tools to improve our quality of life, they often pose a risk and, unfortunately, also produce negative results on a social level. In fact, O'Neil points out how these systems used in an unregulated manner lead to greater discrimination due to their black box status, a concept to which we will return later.

In addition to the benefits of using AI systems, such as process automation, error minimisation, optimisation and efficiency, greater precision and improved decision-making, there are a number of disadvantages that must be taken into account when assessing the suitability of applying AI to a product or service. In recent years, the advancement of computer systems and the successful deployment of technological services have been affected by the scandals caused by the indiscriminate use of AI. According to Narayanan (2019), concerns about AI can be divided into three categories based on its development and use:

- The application of AI in **percepcion issues**, which we find in image recognition software, facial recognition, medical diagnosis through images or deepfakes that are of concern, especially in ethical issues due to their precision, as they currently surpass human capacity.

- The application of AI in **judgement automation**, which assist humans with content recommendations, spam detection or hate speech, and while their development is far from perfect, they continue to improve in trying to mimic human reasoning, and this gives rise to an ethical concern about inevitable errors.

- The application of AI for the **prediction of social phenomena** such as predictive policing, criminal recidivism, employment prediction, etc., which are of ethical concern because of the simplicity of their treatment of social issues and their high degree of inaccuracy in prediction.

AI is a range of many related technologies, some of which are making remarkable progress. However, in many cases the downsides of its use have to do with the exploitation of the wrong AI label in products that are not AI. Narayanan himself warns us that this technology has become "snake oil" and many companies are benefiting from the confusion it generates in the general public. This issue is particularly sensitive if we consider that fields such as public administration, security or justice are carrying out important modifications by adding predictive algorithms to their processes. For this reason, and with the aim of exploring and evaluating the consequences of the current uses and implementations of decision-making algorithms that are affecting society, new non-profit organisations such as ProPublica and AlgorithmWatch have been formed. Thanks to this public work, different cases have come to light in which the risks in the use of AI systems become latent and allow the emergence of ethical AI to be explained.

Perhaps one of the most prominent cases at present is COMPAS, a software currently used in different states in the United States to predict the degree of recidivism and violent recidivism of a prisoner. The risk calculation is established through a questionnaire and is calculated on the basis of the convict's criminal record. However, the programme is not without controversy as the algorithm has a significant bias by miscalculating a higher risk of recidivism if the defendant is black. ProPublica (Larson et al., 2019) compared the recidivism risk categories predicted by the COMPAS tool with defendants' actual recidivism rates in the two years following their rating, and found that the rating correctly predicted an offender's recidivism 61% of the time, but was only correct in its predictions. of violent recidivism 20% of the time.

A similar problem is what we have seen with regard to the distribution of state financial support in Spain. The BOSCO Electricity Social Bonus launched in 2017 by the central government is a software used to determine who can receive financial assistance on their electricity bill. Although the aim was to make it easier for applicants to apply for aid, as well as to streamline the administrative process, the system has received a large number of complaints about the denial of aid to people who met all the requirements, as well as a lack of transparency on the part of the system. Civio, a non-profit organisation, found that BOSCO systematically denied assistance to eligible applicants such as pensioners or widows. At the government's estimate of 2.5 million vulnerable households that would benefit from the social bonus out of a possible 5.5 million, only 1.1 million people have received the aid. Civio thus asked the government for the source code of BOSCO to identify the problem, although the Transparency and Good Governance Committee ended up refusing to share the code, citing possible violation of copyright regulations. In July 2019, Civio filed an administrative complaint claiming that the source code of any system used by the public administration should be made public by default, and this process is still open.

While we have seen great progress in computing in recent years that explains the development of AI, we still have a long way to go before machines reach the same level of thinking as people. We often speak of competence without understanding, so while we cannot ask machines to perform ethical operations, we do need to set ethical parameters in the design of this technology to help curb the implicit risks involved in its use.

1.5. What are the main risks of AI?

In the same *White Paper on AI* (European Commission, 2020), it is noted that the main risks related to the use of AI concern the application of rules designed to protect fundamental rights (including personal data and the protection of privacy and non-discrimination), as well as issues related to security and liability (Cortés et al., 2021). In that sense, it is not surprising that the increased use of technologies that include AI systems has opened a broad debate on the impact and consequences of their use.

Current applications of both AI and automated decision-making algorithms have been applied in fields ranging from surveillance, predictive policing and the health sector, entertainment, and now also have an increasing presence in different financial aid systems, as well as in administrative procedures. More generally, AI has been included in the analysis of large masses of aggregated and anonymised population data, in order to obtain real-time information on crowd behaviour, predict areas of risk and model public policy interventions. We should add that the current health crisis caused by SARS-CoV-2 has caused an acceleration in the use of AI systems and a rapid advance in both facial recognition devices and in geolocation and population control systems for mitigating the virus, as well as for controlling its infectivity.

As AI has become more and more embedded in our daily lives, we have also realised that AI systems can maintain and even amplify different negative biases towards different groups of people, such as women, older people, people with disabilities and also towards minority ethnicities, rationalised groups or other vulnerable groups (Kraemer et al., 2011; Mittelstadt et al., 2016). As a consequence, one of the most pressing questions, especially in the context of machine learning, is how to avoid bias in AI systems (Daniels et al., 2019). Considering that one of the main goals of AI systems is to achieve "greater efficiency, accuracy, scale and speed of AI to make decisions and find the best answers" (World Economic Forum, 2018: 6), the existence of bias in the use of AI can not only undermine this seemingly positive situation in several ways, but also generate a lack of confidence in this technology, especially among those most affected by the biases.

Several studies have highlighted the existence of different biases, mainly through the content and use of websites (Baeza-Yates, 2018), such as those related to the hiring process (Dastin, 2018) and particularly in the sense that certain minority population

groups that have been historically racialised are offered only certain types of jobs (Sweeney, 2013). Another reported racial bias relates to some of the decisions about the creditworthiness of loan applicants (Ludwig, 2015). This is in addition to the notorious racial bias in automated decisions about the release of parolees (Angwin et al., 2016), the bias related to predicting criminal activity in urban areas (O'Neil, 2016), or the bias in facial recognition systems that tend not to identify people with darker skin colour as accurately (Buolamwini and Gebru, 2018). Another important bias stems from AI systems that aim to identify a person's sexual orientation (Wang and Kosinski, 2018).

Not surprisingly, the debate on the main ethical concerns arising from the use of AI models is still open and, for this reason, we should at least take the following aspects into account:

- **Explainability and transparency,** which is given by inscrutable or inconclusive results as well as biased results that negatively affect a sector of the population and accentuate social injustices.

- Security and privacy, the effects that this technology is having as surveillance systems can end up violating the rights of users by threatening the right to be forgotten, not to be exposed or to control information.

- Responsibility, which we refer to within user data management, responsibility and accountability management.

- Well-being, when we talk about the alignment of AI with human values and human rights, we are particularly concerned about the sustainable development of the technology and the environmental impact it produces.

- Autonomy, the maintenance of people's autonomy must prevail over technological development.

- **Solidarity** with more vulnerable communities and the treatment of individuals in a dignified manner.

Among the concerns arising from the use of AI, we can mainly recognise two major risks whose mitigation is essential when building an ethical AI: on the one hand, the abuse of Big Data and algorithmic biases and, on the other hand, the existence of AI black boxes.

I. Big Data and biases in AI

It is currently assumed that the more data used the better. However, this can lead to the use of unrepresentative or biased data so that algorithms do not perform well, commonly known as Garbage In Garbage Out. The use of contaminated, inaccurate or incomplete databases leads to biases and this is one of the main problems in AI development. Many AI systems, such as those containing supervised machine learning, rely on large amounts of data to function properly. In this context of massive data use we can recognise at least three reasons for bias: (1) data bias, (2) computational or algorithmic bias and (3) outcome or selection bias (Springer et al., 2018).

The first problem is that the use of data containing implicit or explicit imbalances not only reinforces a distortion in the data but also affects any decision making, making the bias potentially systematic. The second problem is that an AI system can suffer from algorithmic bias due to the implicit or explicit biases of the developer. This is largely because the design of a programme is based on the developer's understanding of other people's normative and non-normative values. It is therefore important to include users and stakeholders affected by the development process (Dobbe et al., 2018). The third problem relates to outcome or selection bias that is often associated with the use of historical records but also relates to the systematic selection of groups of people and places that become linked to particular outcomes. For example, in predicting criminal activity in particular urban areas, an AI system may end up assigning more police to a particular area because of historical records and a selection by the police command to police some areas much more than others. This logic results not only in more crime being reported in one area but also in more police being assigned to a certain area due to the biased results of the AI system. This seems to be a recurring problem despite the fact that other urban areas may have a similar or even higher number of crimes, many of which would go unreported due to the lack of policing through AI systems (O'Neil, 2016).

We can also define biases in AI in the way data or algorithmic responses reflect the implicit values of humans. Although there are a large number of cognitive biases (Lu, 2020), when referring to Big Data and AI we must distinguish between two main biases: (1) explicit bias (conscious or cognitive bias) and (2) implicit bias (unconscious bias). The first refers to biases for or against one thing, person or group (usually in comparison to another), and are the result of past experiences, which are often shaped by the culture of the place and also by our upbringing. The second is present in our brain, but appears

independently of any conscious cognition.

Cases like the one we have seen with COMPAS accentuate the problem of the use of biased databases. But this is not the only case where biases can occur. Training an algorithm with biased databases or not using a complete database can also lead to biases. This problem is particularly acute in facial recognition software, which fails to recognise women or black people. Joy Buolamwini of the Algorithmic Justice League (AJL) has conducted research to understand and expose the current biases in leading face recognition systems. The *Gender Shader* project published by MIT (Buolamwini, 2018) evaluates the accuracy of different AI systems for gender classification. Through a database of 1270 images of people from African and European countries, classified under gender and skin type labels, they were exposed to facial recognition algorithms from IBM, Microsoft and Face++. The alarming result showed a margin of error of up to almost 35% higher when identifying dark-skinned women versus lighter-skinned men, highlighting the latent prejudices in society.

While eliminating bias completely is an almost impossible task because it would mean eliminating bias in society, different actions can be taken to mitigate its occurrence and reduce the biases that occur in AI. This is why different partial solutions are being promoted:

- From Big Data to Good Data, which means using massive (or not) data that is as representative as possible.

- Promote diverse and inclusive AI including under-represented communities to make actors and data visible.

- Understand and measure biases to include AI solutions against discrimination or simply not to use AI in some cases.

- Generate new training databases comprising a wider and more inclusive sample of data.

- Promoting the creation of teams that are diverse in age, race, culture or gender can benefit the creation of algorithmic models from different perspectives that help answer questions to mitigate the generation of bias.

When auditing an algorithm for bias, two types of factors must be taken into account: equality of outcome and equality of opportunity. For example, if we are talking about

a loan management AI, we can say that the outcome is equal if people from any city get the loan in the same proportion. Similarly, when we refer to equality of opportunity, we mean that those selected receive the same interest rate regardless of the city they come from.

Obviously, apart from taking technological measures to avoid the introduction of biases in data and AI systems, one would expect that the reduction of social inequalities between different groups of people would also lead to a reduction of biases associated with biases, and the latter appears to be essential to avoid situations of discrimination and, ultimately, bias in the data and results provided by some AI systems. However, most AI developers, as well as academics working in the field of technology and computational social sciences, believe that we will never be able to design a totally unbiased system, not least because AI systems, especially machine learning systems, are designed to discriminate, to differentiate, between things like people, images or documents. Nonetheless, there are some types of discrimination that are considered socially undesirable and there are certainly some patterns that should not be used or replicated, as they might be related to legal concepts of discrimination, such as avoiding the direct (or indirect) use of protected characteristics, such as sex or gender, ethnicity or national origin, or disability. At other times, the characteristics of the data used in some AI systems may amplify geographical or social inequalities, and lead to the failure of many social policies.

In this sense, if we are not vigilant, we may find that some AI systems cause inequalities between social groups to be amplified and even more enduring. Clearly, if this is not to happen, a better understanding of the social in the data used for AI systems is required. For example, as some authors point out (Joyce et al., 2021), some AI practitioners may be unaware that data about X (such as postcodes, health records, road locations) is also at the same time data about Y (such as sex or gender, ethnicity, socio-economic status), and may think of data about X as neutral data that applies to all people equally rather than understanding that postcodes very often provide information about discrimination, inequality and social segregation. Certainly, indirect discrimination, where variables that we did not think were sensitive to the proxy, such as sex or gender or ethnicity, pose a big challenge.

Thus, if we do not take these basic issues of social structure into consideration, when identifying correlations between vulnerable groups and life chances, for example through postcodes, it may be that people using AI systems will accept these correlations as causation and use them to make decisions about present and future social interventions. This lack of understanding of the social through AI is a handicap both in

the collection and processing of massive data and in its use through AI systems. Thus, given that most algorithms learn from the massive data collected in society, this data cannot be independent of society, and neither can it be independent of its designers (Smith, 2019).

In this sense, authors such as Floridi (2020) argue that the future of AI lies in data acquisition itself. This technology has evolved so much in recent years that we are moving from an emphasis on Big Data to an emphasis on data quality. And therefore, training algorithms based on smaller, better selected and more reliable data sets will increasingly move us away from the Big Data model towards small data. Floridi also believes that, if this quality is a determining factor, then we must take into account the origin of the data and, in particular, the use of historical data, which can be problematic, as these data are often inaccurate, contain biases, come from unreliable sources or their access is restricted due to privacy issues. Therefore, using synthetic data, i.e. data generated by AI itself, could be a breakthrough in the creation of databases for training algorithms, with more reliable, less biased, easily duplicable, reusable and freely shareable data.

But it should be added that it is not only a question of data quality but also of the business model. In this regard, authors such as Carissa Véliz (2020) argue that we have more than enough evidence to affirm that the current data economy is a toxic business model, among other things because over the last two decades we have allowed many corporations access to our personal data for commercial and marketing purposes, and to this end the use of AI has been crucial for both its collection and massive analysis. In this sense, opinions are increasingly divided between those who see the massive use of personal data as a danger, between those who believe that we should employ AI even though it is not always very clear whether it benefits us individually and collectively, and between those who think we should alert the more optimistic about the risks and injustices that artificial intelligence is causing (Salas, 2019).

II. Black boxes in AI

As discussed above, many of today's AI systems are based on a connectionist approach, be it computer vision, natural language processing or operations research, among others. This approach, which is very successful in learning from statistically correlated data, also has a problem in that it relies more on our intuition than our understanding to explain a general idea of why they work. It is for this reason that we perceive them as black box systems, namely a problem of transparency, explainability and, ultimately, opacity, which is why many current AI systems are referred to as black boxes (Rudin, 2006). As we have already noted, this black box approach is very different from what we had previously when formal logical frameworks were the norm in symbolic AI, i.e. when learned rules used to be present in a human-readable format. With connectionist AI, it is often difficult to understand or trace back the process by which these systems reach certain solutions or predictions. For this reason, many authors discussing the ethics of AI propose explainability as a basic ethical criterion, including among others, for the acceptability of AI decision-making (Floridi et al., 2018).

The opacity of AI decision-making can be of different types. On the one hand, we witness how not even experts can understand the functioning of so-called black boxes (Wachter, Mittelstadt and Russell, 2018) while, on the other hand, there is also opacity with the people who are affected by their use due to the trade secrecy of many AI systems. Thus, the fact that one cannot or will not disclose how a certain automated decision is taken results in an undesirable opacity situation in the development of current AI systems. The normalisation of this situation is problematic, especially in democratic systems where transparency is a fundamental principle, and therefore the implications of this inability to understand the decision-making process are profound at the individual and collective level. Such opacity appears as an affront to a person's dignity and autonomy when decisions about important aspects of their lives are made by AI systems but we cannot explain why the systems adopted certain solutions or decisions. Well-documented and accessible algorithms can provide information about the automated decision-making process, thus increasing the transparency and accountability of the AI system. However, there is a tension with making algorithms completely open to increase transparency and a possible breach of confidentiality when AI systems are trained on personal data. In that case, when algorithms are used to manage complex systems but require more openness or transparency, there is a tendency to favour so-called decentralised systems, where there is no single place where data is stored or verified, such as in a *blockchain*.

These can provide many benefits, such as independence from political control, public verifiability and security against certain types of interference or attack (Dinh and Thai, 2018). In other cases, particularly around machine learning, technologies are being developed to "open" so-called black boxes (Ribeiro et al., 2016). Further, it should also be considered that the fact that there is not full transparency is not always a problem if the solution is positive or beneficial. For example, Robins (2019) points out that, if an AI system could reliably detect or predict some type of cancer in a way that we cannot explain or understand, the value of knowing the information would outweigh any concerns about not knowing how the AI system would have arrived at this conclusion. Thus, this point made by Scott Robbins is relevant and highlights that a strict requirement of explainability could impede some technological advances in AI and its potential benefits.

Even so, it is believed that the growing prominence of algorithmic decision-making without accountability for reasons of opacity or the generalisation of black boxes may become a threat to our democratic processes. For this reason, some authors (including Scott Robbins himself) point to the relevance of distinguishing between contexts of application, especially between those in which the procedure behind a decision is important in itself and those in which only the quality of the outcome matters (Danaher and Robbins, 2020). Within this context, progress in terms of establishing causal relationships appears to be fundamental. In other words, even though some AI systems can determine that X is the cause of Y, this does not mean that X is the only cause of Y. In that sense, Bathaee (2018) proposes two partial solutions: (1) regulate the degree of transparency that an AI must expose; and (2) impose strict liability for the harms caused by the AI. Yavar Bathaee believes that we may be able to reverse this black box effect if we can adjust the tests of intentionality and causality to the level of transparency of each AI system, especially when it makes autonomous decisions. Meanwhile, otthers such as Rudin (2019) believe that creating methods to explain black box models is a strategy that will only continue to perpetuate malpractice and therefore the only viable solution is to create models that are inherently interpretable. In fact, according to Cynthia Rudin, the GDPR model and other AI regulatory initiatives undertaken by the European Union are governed under the law of explainability rather than promoting the creation of interpretable models. This could have implications if we want to avoid false negatives (e.g. an automated car not stopping when it should) and false positives (e.g. investigating an innocent family for child abuse). For this reason, it is proposed that black box models should not be applied when the situation calls for a high-risk decision. Thus, opacity, which is a feature used for intellectual property protection, would conflict with the goals, assumptions and values "embedded" under AI system designs, especially when they have an impact on the public or societal domain.

Moving towards legally transparent AI systems and citizens' meaningful understanding of the decisions that concern them will not be an easy task. However, the fact that there is growing research (Hitzler et al., 2020) to transform a symbolic system into a connectionist one with the aim of realising flexible symbolic learning capable of explaining neural networks after big data training is a sign of the current limitations and future opportunities of AI. Clearly, regardless of technical advances, reliable and empowered intermediaries are needed in order to be able to inform and communicate about the state of the art in AI-related technologies, applications and concerns so that, ultimately, different social actors and individuals do not have to seek costly and time-consuming legal recourse. That is why a central registry of AI systems with open source and/or metadata about them would improve the democratic process and unpack a non-negligible part of the current black boxes.

1.6. The social perception of AI

Society's perception of the implementation of AI systems undoubtedly has an impact on the progress and development of this type of technology. It is for this reason that analysing the degree of trust or social perception regarding the progress of AI has become an important task both for governments and for the gradual civic involvement in its development. Currently, the social perception of AI is very much shaped by public interventions by experts or influencers, who are major drivers of AI risk perception. In fact, since the phrase "artificial intelligence" has been popularised through media and social platforms, interventions by people considered experts have also increased, and some of them have had a great echo. For example, at the end of 2014, an interview with Stephen Hawking on the BBC about the potential threat of developing super artificial intelligence went viral, especially when he stated that "the development of full artificial intelligence could mean the end of the human race" (Cellar-Jones, 2014). In fact, his interview accounted for 14.6% of all AI-related posts and 46.5% of all risk perception posts about AI at the end of 2014 (Neri and Cozman, 2020).

But along with this abstract representation of public awareness of AI through experts, there are also specific surveys for the population as a whole that allow us to capture whether AI is perceived as a benefit or not, especially for the future, and whether this perception is unequal depending on the geographical region where we live. On this issue and from an international perspective, the *World Risk Poll* published in 2019 shows that the population's perception of AI in 20 years' time is very uneven depending on the world region (Neudert et al., 2020). For example, the perception of people living in North America and Western Europe tends to be seen more as detrimental than beneficial, while in South and East Asian countries they are rather optimistic and see the development of AI as beneficial for the future. These differences may seem surprising if we take other studies such as that of Fast and Horvitz (2017), which looks at opinions expressed about AI in the New York Times over a 30-year period, and in which news and discussions mainly through experts, have generally been more optimistic than pessimistic. However, the same study also shows that there are news or concerns about AI, such as ethical considerations, especially a certain loss of control or the possible negative impact of AI on work, which have a very significant impact on the social perception of AI.

Beyond North America, according to the same World Risk Poll, it is in Europe that the view of AI is rather pessimistic, with 43% of respondents believing it will be detrimental

compared to 38% who believe it will be beneficial. This scepticism seems to be more pronounced in Mediterranean countries such as Spain, Portugal and Greece, and also in Belgium, where 50% of respondents are pessimistic about the development of AI in the next 20 years. However, this assessment is not only exclusive to Europe, but also found in other global regions such as Latin America and the Caribbean, as well as in North America, where 49% and 47% respectively believe that AI will harm people in the next two decades. This is certainly in contrast to what the same survey shows about East Asian countries such as South Korea, Japan and especially China, where around 59% believe that AI will be mainly beneficial and only 9% believe that it will be the opposite in the coming years.

Other studies on perception in the immediate context (Lozano et al., 2021) show other relevant aspects, such as the existence of statistically significant differences between men and women with respect to attitudes towards AI, which the same study suggests may have to do with the lower presence of women in AI research and development. Also that there is a negative attitude on the part of some people if they are not interested in scientific discoveries and technological developments or if AI and robots are not seen as useful for the development of their work. On the same topic and context, another study by COTEC (2021) notes that the most vulnerable groups are more pessimistic about their ability to compete in an automated labour market after the pandemic, and although in the same study respondents believe that technological change creates more jobs than it destroys, they also consider that these changes increase social inequality.

This risk analysis, which examines the concept of risk as the statistical expectation of events and sometimes the magnitude of their consequences (Freudenburg, 1988), is often questioned because it ignores important dimensions such as epistemological, sociological and subjective. However, it is also considered useful as a general thermometer and as a way to counteract perceived risk or unintended side effects, in this case of AI development and implementation. In this sense, the possible side-effects of AI range from the more plausible ones, such as the concern and, at the same time, mitigation of biases in AI, to the less plausible ones such as a general AI that is primarily negative or maleficent. Therefore, these exercises in capturing societal perceptions of AI serve, among others, to carry out AI governance actions, such as the creation of institutions to help us improve the current state of AI. Clearly, the less we know about an activity, the more likely it is that the evaluation will become an exercise or a matter of trust, also for the future.

En este sentido, uno de los temas recurrentes que se presenta frecuentemente en eBut while the prospect of massive redundancy and a jobless future has taken hold in the

discourse of the so-called “Future of Work”, some mainly expert opinions believe that technology can also become a catalyst for further work by creating new opportunities for commodification and that, as in the past, it could increase the number of jobs (Huws, 2014). The main thrust of this position is the creation of new value chains, especially the marketing opportunities opened up by the internet, which allow companies to sell more output, while also supporting employment. A second critique concerns capitalism’s ability to reproduce and indeed expand work, albeit often on inferior terms and conditions for working people (Spencer, 2018). It is in this context of competition and reflexive critique that the expert opinion is necessary as it allows, among others, to connect this second critique with the first in the sense that, the use of technology is not only necessary to expand marketing opportunities, but also to maintain work and consumption. It is from this perspective that one can imagine, for example, a future where low-wage and low-productivity work proliferates, and where weak bargaining power also appears as a constraint to working less. For many people this is already the present, where technology is used to expand work opportunities in a way that harms the interests of workers in low-paid, unregulated and insecure work (Friedman, 2014).

But perceptions are not only about the future of work, another major risk posed by AI is its use for weapons. Thus, it is to be expected that the more powerful a technology becomes, the more it can be used for nefarious reasons such as warfare. This could occur if AI systems are used maliciously, not only to produce robots to replace human soldiers, but also for the manufacture of lethal autonomous weapons. On this last point, it should be stressed that the development of autonomous weapons is currently concentrated in countries and regions (the United States, China, Russia, South Korea and the European Union) that have the resources to invest in advanced robotics and AI research. But Moore’s law and falling production costs will soon allow many states and non-state actors to acquire autonomous weapons, which may also erode fundamental norms of international law against the use of force (Haner and Garcia, 2019).

Proof of this is that, since 2013, the debate has moved up the UN arms control agenda, and is an issue for the Human Rights Council, the General Assembly and the Convention on Conventional Weapons. Until a resolution is reached, a 2020 survey on the same topic (with 19,000 people from 28 different countries) indicated that 62% of people oppose the use of lethal autonomous weapons systems, while 21% support them, and 17% are unsure (Ipsos, 2020). Notably, the five countries most active in the development of autonomous weapons, more than half of respondents opposed: Russia (58%), the UK (56%), the US (55%), China (53%) and Israel (53%). However, the results of this global survey do not seem to have much effect as the development of lethal autonomous weapons remains in an accelerating phase in many countries, with millions of dollars

being spent each year, and with almost no public debate. But we do not only have to worry about adversaries, we also have to worry about uncontrolled AI, especially in the military domain, and this does not mean that AI can become "evil", rather we can imagine an AI system that makes decisions autonomously, but with terrible unintended consequences. For this reason, the risk of simply "switching off" is often raised, as this can lead to massive destruction due to the inability of AI systems not only to lack ingenuity, but also to not enjoy a dynamic understanding of all contexts of action.

Undoubtedly, every misstep in AI has consequences for society's perception of AI. From opaque automated decisions, to biases in algorithms and their societal impact on the future of work and weapons use, they can create an unfavourable climate for the adoption of some AI systems. They can also have an impact on the AI narrative, creating false expectations and perceptions that are difficult to reverse. Indeed, it can be argued that as the technologies themselves have developed, from automata to robots and from cybernetics to machine learning today, so too have the associated hopes and fears. In this respect, experts generally exhibit three different positions according to some studies (Neri and Cozman, 2020): they can be antagonistic experts, neutral experts, and enthusiastic experts. The perception of the former is that there are obstacles that are difficult to overcome in order to achieve full-fledged AI at the human level. Meanwhile, neutral or pragmatic experts would be those who consider that it is difficult to represent what the real challenges are in developing full human-level AI, even though there is some conviction that it can be achieved. Finally, the perception of the enthusiastic experts would be that AI is all a matter of time, and that its development will bring about profound change, which may be positive (among the optimists) or negative (among the pessimists). It is worth remembering, however, that just as the public is influenced by perception through emotion and affect in a simple and sophisticated way, so too are experts. Just as worldviews, ideologies and values influence the general public, so do experts.

1.7. What is the institutional response?

While AI technologies have the potential for social and economic development, they also present complex challenges in both the public and private spheres, as well as significant concerns about the automation of prejudice, stereotypes and discrimination. As noted above, different strategies have been developed to address the challenges posed by the adoption of AI systems and also to minimise some of the adverse effects they may have. In general, the global response has been to adopt recommendations and also to create different lines of action and regulatory measures that affect the development and implementation of AI in each country. In this regard, during 2017, a large number of national AI strategies began to emerge, with Canada, Japan, Singapore, China, the United Arab Emirates and Finland being the first group of countries to make public their AI strategy, mainly focused on research and development as well as on attracting talent (Dutton, 2018). Thus, economic growth appears as a common thread in most strategies and, at the same time, their potential for application in the public sector is seen. But it is also clear that there is a growing concern for ethical values in AI, and this has also been reflected to a large extent in the national strategies of some countries such as New Zealand, Singapore, the UK or Sweden.

For example, New Zealand examines how AI will affect law and ethics in areas such as fairness, transparency and accountability, while Sweden proposes to increase basic and applied research in AI and develop a legal framework to ensure sustainable AI development, with an emphasis on AI applications being ethical, safe, reliable and transparent. Also in Northern Europe, the ministries responsible for digital development in Denmark, Estonia, Finland, the Faroe Islands, Iceland, Latvia, Lithuania, Norway, Sweden and the Aland Islands have formed a Nordic-Baltic alliance to pool resources and develop standards, principles and values to develop ethical and transparent AI. In this context, it is worth noting that an analysis and visualisation of the frequency of concepts appearing in more than 450 documents related to AI strategies by the Council of Europe[5], coming from national authorities, the private sector, international organisations or multi-stakeholder initiatives, shows that two concepts stand out as the most important above all others: human rights and privacy. Undoubtedly, these concepts reflect how, apart from the economic growth implied by the adoption of AI, the ethical and social considerations of its implementation are also a common

[5] Council of Europe (2021) AI Initiatives. Data Visualisation of AI Initiatives. Strasbourg: Council of Europe Portal, https://www.coe.int/en/web/artificial-intelligence/national-initiatives (accessed 27 August 2021).

denominator in most strategies or working documents.

In 2018, the European Commission published the report *Artificial Intelligence: A European Perpective* (Annoni et al., 2018), a document describing the EU's approach to AI. This document describes several objectives such as increasing the EU's technological and industrial capacity and the adoption of AI by the public and private sectors, as well as preparing citizens for the socio-economic changes brought about by AI and creating an appropriate ethical and legal framework. Alongside these measures, in April 2019, the High-Level Expert Group on AI presented the report *Ethics Guidelines for Trustworthy Artificial Intelligence* (European Commission, 2019). The main objective of this document is to promote the development of trustworthy AI, providing a series of recommendations to all actors involved in the process of design, development, implementation and use of AI. Under this report it is defined that AI must be a tool that favours the common good, individual and social well-being and human prosperity, as well as favouring progress, prosperity and innovation. Likewise, the reliability of AI systems rests on three pillars: it must be lawful, ethical and robust. At the same time, this document sets out guidelines for ethical and robust AI through four principles:

1. Respect human autonomy without unjustifiably conditioning or coercing human beings.
2. Prevent harm and damage that this technology may cause.
3. Promote a fair and equitable distribution of benefits and costs, as well as equal opportunities in access to education, goods, services and technologies.
4. Promoting the explainability of decision-making.

Based on these principles, the European Commission foresees a series of actions to be taken to materialise the recognition of this trusted AI such as: support for human action and the principle of user autonomy, protection of vulnerabilities of AI systems, data and underlying IT structures, privacy and data management as well as the right to privacy of users, ensuring traceability, explainability and transparency of all elements of AI.

In this regard, the so-called Ad hoc Committee on Artificial Intelligence (CAHAI) appointed by the European Commission or the Organisation for Economic Co-operation and Development (OECD) expert group on AI in Society are two figures created to advance the deployment of trusted, people-centred AI. In fact, and almost simultaneously, the

OECD in May 2019, recognising the rapid development and deployment of AI and the need for a stable policy environment that promotes a human-centred and democratic values-based approach to AI, adopts five complementary principles formulated by its Digital Economy Policy Committee (OECD, 2019):

1. **Inclusive growth, sustainable development and well-being**
   Stakeholders must proactively engage in the responsible stewardship of trusted AI in pursuit of beneficial outcomes for people and the planet, such as increasing human capabilities and enhancing creativity, promoting the inclusion of under-represented populations, reducing economic, social, gender and other inequalities, and protecting natural environments.

2. **Human-centred values and equity**
   Stakeholders must respect the rule of law, human rights and democratic values throughout the lifecycle of the AI system, including internationally recognised principles of freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, equity, social justice and labour rights. This means that AI actors must implement mechanisms and safeguards, such as human determinability, that are appropriate to the context and in line with the state of the art.

3. **Transparency and explainability**
   Stakeholders should commit to transparency and responsible disclosure with respect to AI systems. In this regard, they should provide meaningful information, appropriate to the context and consistent with the state of the art, including a general understanding of AI systems, their interactions and that affected persons understand the outcome and, in turn, can challenge it based on simple and easy-to-understand information about the factors and logic that formed the basis for the prediction, recommendation or decision.

4. **Robustness, safety and security**

   AI systems must be robust and secure throughout their lifecycle so that, under normal use, they function properly and do not present unreasonable security risks. AI actors must also ensure traceability, including in relation to data sets, processes and decisions made during the AI system lifecycle, and apply a systematic risk management approach to each phase of the AI system lifecycle.

5. **Responsibility**

   AI actors should be responsible for the proper functioning of AI systems and for the respect of the above principles taking into account their roles, the context and in accordance with the state of the art.

Since then, UNESCO (2019) has also led a multidisciplinary, multicultural and pluralistic effort to produce a first draft of recommendations, produced in November 2019 for its General Conference and intended as an international normative instrument on the ethics of AI. In UNESCO's recommendations, special attention is given to the ethical implications of AI in relation to UNESCO's core domains (education, science, culture, and communication and information), as follows:

1. **Education**

   Digital societies require new educational practices, the need for ethical reflection, critical thinking, responsible design practices and new competences given the implications for the labour market and employability.

2. **Science**

   AI technologies bring new research capabilities in a broad sense and include academic fields from the natural sciences and medical sciences to the social sciences and humanities. They then have implications for our concepts of scientific understanding and explanation, and have the capacity to create a new basis for decision-making.

3. **Identity and cultural diversity**

   AI technologies can enrich culture and creativity, but can also lead to a greater concentration of the supply of cultural products, content, data, markets and embodiment in the hands of a few actors, with possible negative implications for diversity and pluralism of languages, media, culture, expressions, participation and equality.

4. **Communication and information**

   AI technologies play an increasingly important role in the processing, structuring and provision of information, including automated journalism, algorithmic news provision and content moderation in social media and search engines. This raises questions related to access to information, misinformation, the emergence of new social narratives, freedom of expression, and so on.

While deploying a global strategy to foster understanding of the potential impacts of AI in different domains as well as synergies and cooperation in AI development and implementation, it is clear that supranational actions such as that of EU Member States together with specific national AI strategies within the EU have served to strengthen the competitiveness of some regions such as the EU in global AI. While in the global race for AI, Europe competes with competing visions such as the so-called 'AI for profit' (United States of America) and 'AI for control' (China), there is an increasingly entrenched vision within European institutions that postulates Europe to adopt 'AI for society', a human-centred approach in which AI systems are safe and ethical by design, as a hallmark of European development in the field of AI (Annoni et al., 2018). In this regard, as stated in the recent European Commission and OECD *AI Watch National Strategies on Artificial Intelligence* report (Van Roy et al., 2021), although different European countries' approaches to AI differ in some strategic priorities, budget allocations and implementation timing, it can be said that all Member States have common objectives to support the adoption and development of AI taking into consideration ethical and societal concerns.

But it is not only the EU Member States that have ambitious plans for the development and promotion of AI with a focus on people and that respects human rights and democratic values. The *Artificial Intelligence Strategy of Catalonia* (CATALONIA.AI, 2020), promoted by the Generalitat de Catalunya and coordinated by the Department

de la Vicepresidència i de Polítiques Digitals i Territori, was launched in February 2020. There are four implementation pillars in this strategy: a pillar of collaborative research through the AIRA alliance (Artificial Intelligence Research Alliance) for institutes and research centres that are benchmarks in AI to develop a coordinated strategy; a pillar for the valorisation of knowledge and innovation through the CIDAI centre (Centre of Innovation for Data tech and Artificial Intelligence); a pillar on ethical considerations and social impact through the OEIAC observatory (Catalan Observatory for Ethics in Artificial Intelligence); and a pillar for entrepreneurship in the AI sector in Catalonia. In order to develop these four pillars, the CATALONIA.AI strategy has a multi-sectoral plan that focuses on six axes:

1. **Ecosystem**
   To foster a governance model that will lead to the development of a coordinated and globally connected AI ecosystem

2. **Research and innovation**
   To boost research and innovation through the application of instruments and synergies between the Administration, research centres and user organisations in AI.

3. **Talent**
   To create, attract and retain specialised talent to drive the development of AI solutions and knowledge transfer to society, while empowering citizens and professionals in other sectors.

4. **Infrastructure and data**
   To provide the necessary infrastructures for the development of AI and to facilitate secure access to public and private data

5. **Adoption of AI**
   To promote the incorporation of AI as a driver of innovation in the Administration and in strategic sectors such as agri-food, health, education, the environment, mobility and tourism, among others.

6. **Ethics and society**

To promote the development of ethical AI, which respects the law, is compatible with our social and cultural norms and is people-centred.

The *CATALONIA.AI* strategic plan is aligned with the European AI development objectives and is born with the aim of becoming a benchmark in Southern Europe. This framework uses the values defined in the *Barcelona Declaration*[6] (2017) as well as sharing the CAHAI recommendations on AI contained in the guidelines of the Montreal Declaration (2018) where people, whether developers or users, are at the centre of policy. One of the fundamental objectives of the strategy, in addition to the development of AI in strategic sectors, is to boost international cooperation, establishing an ecosystem that includes public administration, universities and research centres, industry and civil society and that enables the generation of innovative projects, the attraction of investment and the use of international programmes such as Horizon Europe. With regard to the development and promotion of innovation and research, the Centre of Innovation for Data tech and Artificial Intelligence (CIDAI) has been created as a centre of excellence to accelerate the adoption of technologies in the application of AI. The *CATALONIA.AI* strategic plan has a special concern for the adoption of ethical AI based on people's fundamental rights, including our social and cultural values and ethical principles of autonomy, justice and explainability. This is why the plan includes the creation of the Catalan Observatory for Ethics in Artificial Intelligence (OEIAC), whose main objective is to study the ethical, social and legal consequences as well as the risks and opportunities of the implementation of Artificial Intelligence in everyday life in Catalonia. The OEIAC seeks to have a fully transversal perspective, that is, to take into consideration the presence of the four key pillars in any innovative process: knowledge, public administration, business fabric and citizenship.

At the state level, in December 2020, the Government of Spain also pushed for the creation of a major strategy to strengthen the implementation and development of AI in Spain, known as the *National Strategy for Artificial Intelligence* (ENIA, 2020). This strategy takes a multidisciplinary approach to address economic, social, environmental, public management and governance challenges, and includes perspectives for a wide range of sectors and disciplines. It seeks to boost the growth of AI in the Spanish economy in the coming years with AI policies at national level, while aiming for alignment with EU AI policy. In particular, six objectives stand out in the ENIA strategy:

[6] Barcelona Declaration (2017) Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe, IIIA CSIC, https://www.iiia.csic.es/barcelonadeclaration/ (accessed 27 August 2021).

1. Promote the **development of skilled human capital** in AI through the provision of training and education opportunities, the stimulation of talent and the attraction of global talent.

2. Develop **strong scientific excellence** in the field of AI.

3. **Promote the leadership of tools,** technologies and applications for the projection **and use of** the Spanish **language** in AI.

4. Boost the **deployment and use of AI in the public and private sectors**, including also cross-cutting sector activities and major challenges.

5. **Ensure an ethical framework** that outlines individual and collective rights and creates an environment of trust in AI.

6. **Ensure inclusion in the AI-driven economy**, reducing gender gaps and digital divides while supporting the ecological transition and territorial cohesion.

At the local level, in April 2021, Barcelona drew up a municipal strategy on algorithms and data with an ethical and social perpective, which proposes the development of AI and emerging technologies taking into account three essential aspects (Comissionat d'Innovació Digital et al. , 2021):

1. **Maintain and increase the democratic monitoring** of AI by the public institutions and their citizens.

2. Ensure through **transparency and auditability** that algorithmic models and the databases from which they are applied follow **human rights** and public interest **criteria**.

3. Clarify the **liability regime** for any damage or loss that may arise from the creation and use of AI-based solutions not only by government but also by companies and developers.

As of June 2021, 20 Member States and Norway had already adopted national AI strategies, while 7 EU Member States were in the final drafting phase and planned to publish their strategy in the coming months. The EC-OECD database of national AI policies contains multiple national AI strategies and AI-related policy initiatives from more than 60 countries[7] , which have five AI policies in common (JRC and OECD, 2021):

1. **Human capital**
   Policies to encourage the educational development of people in the use and development of AI solutions, including AI training and identification of future needs.

2. **Market research and innovation**
   Policies to promote AI research and innovation for business growth in the private sector and for greater efficiency of public services.

3. **Networking**
   Policies related to AI mapping, collaboration, dissemination and uptake in the private and/or public sphere to increase international attractiveness and attract foreign AI talent and companies.

4. **Regulation**
   Policies for the development and adoption of ethical principles, legislative reforms and (international) standardisation of AI solutions.

5. **Infrastructure**
   Policies and initiatives to promote the collection, use and sharing of data, and to promote digital and telecommunications infrastructure.

[7] OECD (2021) OECD AI Policy Observatory, https://oecd.ai (accessed 27 August 2021).

Other supranational institutions such as the World Health Organisation (WHO) have also recently published a report presenting guidance on the ethical use of AI in the health sector. Along the same lines as countries or regions of the world, the lack of a general consensus for the ethical use of AI has led to debate among industry players as well as growing concern about the implications of this technology. The WHO report (WHO, 2021), *Ethics and Governance of Artificial Intelligence for Health*, seeks to address similar concerns for national institutions by offering six basic principles for the use of AI:

1. **Protecting autonomy**
   It indicates that decision-making in medicine should be done by humans rather than machines.

2. **Promote human welfare, human safety and the public interest**
   It aims for safety and public interest, stating that AI must not harm people, physically or mentally.

3. **Ensuring transparency, explainability and intelligibility**
   It seeks to improve transparency of the technology not only among developers and regulators, but also for medical professionals and patients affected by it.

4. **Promoting responsibility and accountability**
   It aims that the stakeholders of a given AI product are responsible for ensuring that the technology achieves the expected result and that procedures should be in place to remedy the situation if something goes wrong.

5. **Ensuring inclusion and equity**
   It requires that AI for health be designed for equitable access, with respect to any characteristics protected by human rights codes, such as age, sexual orientation or race. This is especially important as bias in AI remains a prevalent concern.

6. **Promoting responsive and sustainable AI**

   It means that there are minimal negative impacts on the environment. It also indicates that AI products should be continuously assessed during their use.

Importantly, as the public sector increasingly turns to AI systems for decision-making across a range of public services, there is a growing concern and evidence that some of these systems can cause harm and often lack transparency in their implementation. Unsurprisingly, specific regulatory and policy tools are also being used in the public sector, in the hope of ensuring algorithmic accountability. However, as highlighted in the first global study *Algorithmic Accountability for the Public Sector* to analyse the initial phase of algorithmic accountability policies for the public sector (Ada Lovelace Institute, AI Now Institute and Open Government Partnership, 2021)[8], the institutional responses are still emerging and changing rapidly, and vary widely in form and content, from legally binding commitments to high-level principles and voluntary guidelines.

It should therefore be stressed that, despite efforts to develop AI strategies and adopt the recommendations of different institutions such as the EC and the OECD, there is still a long way to go to make the values outlined above tangible, especially in the advancement of AI that systematically incorporates ethical considerations and is socially responsible. The latest report by the Institute of Electrical and Electronic Engineers (IEEE) on *AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection* (Schiff et al., 2021) is along the same lines. In fact, this IEEE report, which analyses 112 policy, strategy and regulatory framework documents from different fields around the world, highlights the consensus reached on the need for social responsibility in AI. At the same time, it underlines that in the private sphere there is also a technical concern about the transparency of algorithms, and that in the public and NGO sphere, apart from accountability, the principle of fairness is considered important, especially through accountability. In addition to these concerns, there is also a certain unease that many strategies, both in terms of their content and how they were created, have too few experts in the field of ethics in particular and social sciences in general and, instead, numerous representatives of industry (Metzinger, 2019). According to the same author and others, this may lead to a situation where guidelines not only fail to account for the potential social impact of AI but also use language that may be socially inaccurate or non-confrontational, which may have the risk of being interpreted as possible 'ethical laundering' (Resseguier and Rodrigues, 2020).

[8] Ada Lovelace Institute, AI Now Institute & Open Government Partnership (2021). Algorithmic Accountability for the Public Sector, https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/ (accessed 27 August 2021).

Another critical aspect that is increasingly being pointed out is the fact that most strategies only take into account Western perspectives on the ethics of AI technology, leaving out non-Western visions such as African and Asian ones. It is for this reason that it is advocated that future versions of such strategies with ethical guidelines should also include non-Western contributions, as AI products and services are clearly global. In this regard, although there has been recent inclusive progress through work on virtue and ethics (Jing and Doorn, 2020) and community and relational ethics (Wareham, 2020), these still remain minority perspectives

1.8. What is the business response?

But if we have looked at the global response and the different institutional strategies for the application of ethical AI, let us now look at what has been or is being the business response to the adoption of AI. Although the initial development of AI in the 1950s was viewed with much scepticism in the business world, with the development of information technologies, the scepticism related to AI has not only diminished, but what exists today is an euphoria in its applicability, both for support in the decision-making process and for the development of solutions that provide a competitive advantage in business. In this sense, AI is very rapidly changing the way information is generated and used for decision-making (Mikalef et al., 2017), and it is also bringing about a revolution in the ways of doing business (Schneider and Leyer, 2019), especially in business and management practices in various sectors that offer increasingly competitive and sustainable products or services (Wirtz and Müller, 2019). Thus, we can say that the combined use of algorithms and a large amount of data, connections and interactions are already part of the standard management of a growing number of business organisations (Schneider and Leyer, 2019).

However, it is becoming increasingly evident that for a better understanding and implementation of AI in the enterprise world, businesses must consider different requirements and expectations of AI, starting with its design. It is for this reason that the IEEE (2021) has recently provided the report *Standard Model Process for Addressing Ethical Concerns during System Design*, with the aim of showcasing a series of standard processes through which companies can consider possible negative impacts associated with the design of AI products or systems. The report known as IEEE 7000-2021 standard contains:

1. A standard systems engineering approach that integrates human and social values into traditional systems engineering and design.
2. Processes for engineers to translate stakeholder values and ethical considerations into system requirements and design practices.
3. And a systematic, transparent and traceable approach to meeting obligations ethically oriented regulators in the design of autonomous AI systems.

The fact that the IEEE makes this proposal largely responds to the fact that the reality of AI business is currently much more advanced than public regulation, which makes companies such as Google, Microsoft, Facebook or SAP, which have AI as one of their business pillars, publish different ethical principles sui generis to show their attention, but not necessarily for their implementation or ethical consistency in their design. For example, in 2018 Sundar Pichai, CEO of Google, published a letter[9] of good intentions in which he mentions that the company has established a series of concrete standards to ensure that its AI is ethical:

1. Be socially beneficial.
2. Avoid creating or reinforcing unfair biases.
3. Be built and tested for safety.
4. Be accountable to people.
5. Incorporate the design of privacy principles.
6. Maintain high standards of scientific excellence.
7. Be available for uses that are in accordance with these principles.

Google also details a number of areas in which it will not develop or implement AI:

1. **Technologies that cause or are likely to cause harm.** Where there is a risk of harm, they shall only proceed when they consider that the benefits substantially outweigh the risks and shall incorporate appropriate safety restrictions

2. **Weapons** or other technologies whose primary purpose or implementation is to directly cause or facilitate injury to persons.

3. **Technologies that collect or use information for surveillance** in violation of internationally accepted standards.

[9] Pichai, S. (2018) AI at Google: our principles, https://blog.google/technology/ai/ai-principles/ (accessed 27 August 2021).

4. **Technologies whose purpose does not infringe** widely accepted **principles of international law** and human rights.

However, the recent scandals over the publication of a paper critical of bias in training data used in AI language technologies (Bender et al., 2021) and the dismissals, first of Google's ethics co-director Timna Gebru, and then of Margaret Mitchell, the founder of Google's Ethical AI and co-founder of ML Fairness at Google Research, have not only fuelled debate about the divisions and interests that can exist in a company of this size in the design and use of AI systems, but also about academic freedom and team diversity, raising questions about Alphabet's commitment to the ethical considerations of its AI. Another giant in the sector that has included a series of ethical principles as the backbone of its company policy is Microsoft (World Economic Forum, 2021). Through these 6 values, the company defines a working framework on which the rest of the development teams must base themselves and which it enforces through the design of a Responsible AI standard (2019):

1. **Equity,** to develop systems that treat all users in a fair and balanced way, understanding the context of programme use and purpose to suit the development and implementation process.

2. **Reliability and safety**, encompasses the development of products that are robust and capable of safe operations in stressful environments, as well as consideration of the harms that can come from a technology and the ways in which employees can strive to minimise these risks.

3. **Privacy and security,** to protect users' data and privacy and to use them in a secure manner for all stakeholders.

4. **Inclusiveness,** not to be limited to a few privileged communities, so that all communities across the spectrum of humanity must participate in the process of technological development. This inclusion should not only involve building for, but building with diverse stakeholders.

5. **Transparency,** so that the technology is intelligible and explainable, not only to those who are developing the technology but also to those who use it. Stakeholders must be able to interpret and understand what a technology does and why it acts the way it does.

6. **Accountability**, to ensure accountability at multiple levels, including design, development, sales, marketing and use, as well as promoting regulation of technologies where warranted.

Alongside this definition of ethical values, Microsoft also applies some more applied initiatives to mitigate the risk of discriminatory bias in the use of ML, such as its own *FairLearn* system, an open source toolkit for data scientists, developers and researchers to assess and improve fairness in machine learning. Through a set of metrics and a data visualisation dashboard, it provides insight into how user groups may be negatively affected by algorithmic models. It also includes different unfairness mitigation algorithms to apply to the different tasks performed by the AI.

Facebook uses something similar through a project called Fairness Flow, which makes it possible to determine whether an ML algorithm contains biases. However, the project is still in an embryonic state (O'Brien, 2020), and although it has been applied to some algorithms used by the company itself (such as those used for recruitment), it is not clear to what extent this tool will be implemented 100% in other activities. IBM also has its own open source analysis tools, based on the study conducted by Buolamwini and Gebru (2018) on gender bias, called *AI Fairness 360* (AIF360), which allows up to 10 types of bias to be determined through labelled data and can be applied in algorithms of different scales.

In a more collaborative line with the users themselves, we find companies such as Twitter, which for some time now has been applying a policy of AI systems based on Responsible Machine Learning[10], which consists of implementing responsible machine learning (ML) systems that are responsive and driven by the user community itself. According to the company itself, its deployment of Responsible ML consists of 4 pillars of action:

[10] Williams, J., & Chowdhury, R. (2021) Introducing our Responsible Machine Learning Initiative, https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative (accessed 27 August 2021).

1. Take responsibility for algorithmic decisions.
2. Provide equitable and fair outcomes.
3. Be transparent about the decisions taken and their process.
4. Allow agency and algorithmic choice.

This last point is certainly innovative and, according to the company, algorithmic choice will allow users to have more say and control in how Twitter is configured. In addition, the company is conducting in-depth analysis and studies to assess the existence of potential harms in the algorithms they use, of which the following examples stand out:

- A racial and gender bias analysis of its image cropping algorithm.
- An evaluation of online recommendations for different racial groups.
- An analysis of content recommendations taking into account different political ideologies in different countries.

Clearly, these and other actions are a major step forward. However, we know that one of the main problems is that ethical considerations in AI are not always a priority or binding for AI services or products, and this is causing the implementation of different ethical principles such as accountability, transparency, privacy, justice or sustainability to be treated more as an impediment to progress than as an essential value for the generation of trust in AI systems. Indeed, it is noted that economic incentives still easily outweigh commitment to ethical, social or fundamental rights values such as beneficence, non-maleficence, justice and explicability (Taddeo and Floridi, 2018). Indeed, although the application of ethical principles may entail reputational losses in case of misconduct, on the whole, these mechanisms are rather weak and do not represent a major threat (Hagendorff, 2020).

For the same reason, some companies are known to present ethical considerations around their AI business in terms of branding, or as a soft policy to shy away from (almost) all AI regulation (Bietti, 2020) or to suggest to legislators that internal self-governance is sufficient, and that no specific laws are needed to mitigate potential technological risks and eliminate abuse scenarios (Calo, 2017).
In this context, Ortega (2020) insists that we must distinguish between gestures that he calls “constructive”, such as the funding of the AI for Good Institute, EU programmes, etc., and those that justify a certain suspicion that seek to wash the donor’s face

beyond promoting the ethical use of AI, such as the $350 million from Blackstone to MIT or the $27 million from the Knight Foundation to the Artificial Intelligence Ethics and Governance Fund. In this regard, the recent report The Lobby Network (Bank et al., 2021) describes for the first time the network of influence of Big Tech and the "universe" of actors that put pressure on the European Union institutions in their digital economy initiatives, ranging from the giants of Silicon Valley to the contenders in Shenzhen[11]. Another of the open fronts on this issue is the training and education received by the engineers and developers in charge of building AI systems, for which universities such as MIT, Stanford or Carnegie Mellon have already added specific courses in ethics and social sciences in general to their curricula. However, there is still a long way to go for ethics and social studies to form a backbone of AI development and implementation.

In Europe, it seems that the regulatory boom and innovation plans are beginning to bear fruit and this is reflected in European funding programmes such as Horizon 2020, which is set to see a three-fold increase in the number of AI-related projects. An example of this is the AI4EU platform, which offers a methodology for the ethical design and verification of AI applications, as well as an observatory at European level that acts as a clearinghouse for ethical, legal, socio-economic and cultural debates within the European Union[12]. In this sense, we can say that Catalonia, through this and other European projects, has positioned itself as a leading region in AI, with the capacity to attract a high volume of competitive European funds, especially in innovation projects in SMEs and the application of AI to societal challenges. Projects related to machine learning account for most of the applications submitted, followed by projects based on artificial vision and natural language processing. It is also important to highlight the recognition of researchers and companies dedicated to AI ethics, which account for 34% of the Catalan participation (Bigas et al. 2021).

Even so, the percentage of companies engaged in AI in Europe is low and the EU average is around 6%. Ireland (20%) and Malta (15%) stand out as the most advanced countries, followed by the Nordic countries Finland (10%) and Denmark (9%). In Spain, the percentage of companies using AI is slightly higher than the European average, at 7%, mainly dedicated to machine learning, service robots and virtual assistants such as chatbots (Misuraca and Van Noordt, 2020; ONTSI, 2021). In terms of the sectors of activity that make most use of AI systems, we find the travel agencies and tour operator reservations branch of activity, which according to the report of the *National Observatory of Technology and Society Indicators of the use of Artificial Intelligence*

[11] Bank, M., Duffy, F., Leyendecker, V., & Silva, M. (2021). The Lobby Network: Big Tech's Web of Influence in the EU, Brussels and Cologne: Corporate Europe Observatory and LobbyControl, https://corporateeurope.org/en/2021/08/lobby-network-big-techs-web-influence-eu (accessed 27 August 2021).

[12] AI4EU (2021). Ethics: Promoting European ethical, legal, cultural and socio-economic values for AI, https://www.ai4europe.eu/ethics (accessed 27 August 2021).

in Spanish companies[13] are the most likely to use it, followed by information and communication companies and the ICT sector, with 13% adoption of AI, and electricity, transport, retail trade and accommodation companies, with uses of AI systems above 10%. The sectors of activity less prone to AI adoption would be construction, metallurgy and real estate activities, which show an implementation of this type of technology of less than 5% (ONTSI, 2021).

[13] National Observatory of Technology and Society (2021). Indicadores de uso de Inteligencia Artificial en las empresas españolas. Madrid: Ministerio de Asuntos Económicos y Transformación Digital, Secretaria General Técnica, https://www.ontsi.red.es/es/dossier-de-indicadores-pdf/indicadores-uso-inteligenciaartificialempresas-espanolas accessed (27 August 2021).

1.9. How to move towards ethical AI?

We have reviewed the main ethical principles and recommendations that have emerged in recent years taking into account the immediate context (Europe) and also their significance at the global level (e.g. OECD, WHO and UNESCO). To date, more than a hundred strategies and frameworks on ethical AI have been published, all of them aiming to provide information on ethical and social impact issues, as well as to monitor the use and development of AI technologies. As Hagendorff (2020) underlines, in approximately 80% of all strategies and frameworks, issues of responsibility, privacy or fairness appear as central and, at the same time, these are the issues where technical solutions can be or have already been developed.

In this regard, enormous efforts are currently being made to reach consensus and meet ethical goals in the fields of accountability and so-called explainable AI (Mittelstadt et al., 2019), in those issues related to conscious and fair data collection so as not to amplify inequalities and discrimination (Gebru et al., 2018), as well as in the field of privacy (Baron and Musolesi 2017). While several companies already offer tools for mitigating bias and improving their fairness, accountability and transparency through FAT ML or Fairness, *Accountability and Transparency Machine Learning*, and XAI or *Explainable Artificial Intelligence communities* (Veale and Binns, 2017), we should note that their adoption may be different for large and small companies and for different public administrations, although according to the IEEE (2019) any use of AI systems should always take into account universal human values, data agency and technical reliability in any set of principles to guide their design and implementation.

Thus, for both the private and the public sector, ethical and social impact considerations cannot be understood in opposition to the implementation of AI solutions, nor should they be seen as a mere intellectualisation of complex problems that can only be addressed from a technological point of view. It is for this reason that in order to move towards ethical AI in organisations and institutions, approaches must be adopted that, at a minimum, should conform to the following design and evaluation criteria (Wagner, 2018):

1. **External engagement**, which means early engagement with all stakeholders.

2. Provide an independent (not necessarily public) **external monitoring mechanism.**

3. **Ensure transparent decision-making procedures** on why certain decisions are taken.

4. **Develop a stable list of** non-arbitrary **standards** where the selection of certain values and rights can be plausibly justified.

5. **Ensure that ethics is not a substitute** for fundamental rights and human rights.

6. **Provide a clear statement** on the relationship between the commitments made and existing legal or regulatory frameworks, in particular what happens when the two are in conflict.

We know that one of the problems with how ethics can be applied to AI systems is that in their evaluation, primarily from a risk perspective, the subjective judgements of regulators also permeate the processes of risk identification, assessment and evaluation (Redmill, 2002). However, while subjectivity will always exist in the initial assessment of an AI system, it is important that regulators proactively enforce classifications and controls in business and institutional AI use cases to manage the ethical risk that may arise in their respective industries and administrations, as highlighted in the *White Paper* (European Commission, 2020). It is therefore recommended that, as a starting point, all commercial and governmental use of AI should strictly comply with current regulatory proposals as otherwise it places a disproportionate burden on those who are adversely affected by the systems of AI once implementation has taken place.

In this regard, and to ensure that regulators are prepared, some authors (Huang et

al., 2021) recommend cross-disciplinary training to generate knowledge about the intersection of AI technology and domain expertise in each sector. According to the same authors, this interdisciplinary training is distinct from the sum of knowledge of different individuals and enables regulators to fundamentally understand the risks posed by AI at the use case level.

In Europe, this interdisciplinary vision can be found in organisations such as the AI Ethics Impact Group (AIEI Group), led by the VDE Association for Electrical, Electronic and Information Technologies and Bertelsmann Stiftung[14]. The aim of this consortium is to put AI ethics into practice by means of labelling frameworks and specifications that enable transparency and quality comparability of AI systems in the market. To this end, they propose the use of a model called VCIO (Values, Criteria, Indicators, Observables) to make a series of ethical principles or values practicable, comparable and measurable. In this sense, the general rule recommended by the group to deal with ethical IA will be to apply assessment criteria and indicators that can be observable. Following this model, the AIEI Group proposes the creation of an ethical label for AI systems, similar to the EU energy efficiency label used for household appliances. Such a label could not only improve the competitiveness of a given application, it could also incentivise the ethical development of AI beyond the current legal requirements. To assess compliance with ethical principles, AIEI Group proposes the establishment of six key values to serve as a yardstick, namely transparency, accountability, privacy, fairness, trustworthiness and environmental sustainability. In terms of deciding what level of label should be considered ethically acceptable, it should be noted that this may vary depending on the sector where it is applied. An AI system used in an industrial process where it may be subject to a lower level of transparency is not the same as the same system applied to a medical procedure involving the processing of personal data.

This is why the model proposes the use of a risk matrix that allows the classification of systems depending on their degree of implementation. Following the recommendations of the AIEI Group consortium, the general description of how different stakeholders can use this approach is as follows:

[14]AIEI Group (2020). AI Ethics Impact Group: From Principles to Practice – An interdisciplinary framework to operationalise AI ethics, https://www.ai-ethics-impact.org/en (accessed 27 August 2021).

1. If an organisation plans to use an AI system for a specific application, then it must **determine whether the application is ethically sensitive** based on the experience of the HLEG pilots (European Commission, 2019). Thus, if an application is classified as not ethically sensitive, the process ends at this stage (e.g. it might not be necessary in purely industrial applications). However, if there are ethical issues to consider in the application, then the organisation should conduct a full assessment of the context of the application using a risk matrix. In case there is no official regulation or standard for your field of application, then the VCIO approach can be used.

---

2. In both the public and private sector, departments procuring and using AI systems should use an **ethical rating and risk matrix** to provide clear specifications on the AI systems they plan to use. This procedure not only benefits market transparency through an ethical labelling of AI systems, but also allows for filtering and reviewing product catalogues or visiting websites with the desired ethical rating.

---

3. Manufacturers of AI systems can use the risk matrix and **decide whether to market** an AI system only for applications that do not require ethical sensitivity, or also for higher risk classes.

---

4. Regulators can use the combination of the risk matrix and the ethical rating to **specify requirements for different application contexts and to avoid over-regulation** of fields of application that do not pose any significant ethical challenges.

---

5. At the same time, consumers can **use the ethics rating to compare AI products and services** and make informed decisions about what is acceptable to them and/or worth investing in.

At this point, it should be emphasised that the checkbox guidelines should not be the only tools of AI ethics, nor is it a matter of subsuming as many cases as possible under individual principles in an overgeneralising way, as we need to take into account individual

situations and contexts as well as technical specificities in different areas of application (Krafft and Zweig, 2019). Indeed, the incorporation of ethics cannot be taken for granted only by considering solutions of a technical nature, as the management of corporate and institutional culture also counts. Too often, companies or institutions do not use ethical principles in their management and for ethics to penetrate any organisation, we must first incorporate ethical values to build ethical cultures at the corporate level (Grandy and Sliwa, 2017). When the set of shared values is ethically conceivable and credible both inside and outside, then organisational cultures are ethically sound and only need to be revised as new obligations arise (Rothschild, 2016). For this to happen, there are three conditions that facilitate the exploitation of opportunities to promote ethics in any organisation, including accountability to society, moral autonomy and a climate of mutual trust, as well as ethical deliberation itself (Martinez et al., 2021).

For this reason, without the collaboration of society as part of the decision-making models, the measures adopted by the institutions only cover part of the ethical principles. This also entails promoting education on people's digital rights, as this is an essential part of ensuring the appropriateness of ethical AI. Thus, it is essential to make users themselves co-responsible for the consumption of smart technologies, as it is they who will be able to decide whether or not to use them and, therefore, help define the market and the technological advances that can be developed. But what is clear is that for any AI system to be ethically conceivable and credible both inside and outside an organisation, it will require an interdisciplinary and multi-sectoral approach. For this reason, the involvement of different stakeholders is crucial, from a technical, political and civil society perspective, in order to define and adopt a customised approach to address different ethical considerations in AI systems.

Certainly, both the public and private sectors appear to have joined forces to respond quickly to the risks of AI use. But this dynamic needs to be accompanied by greater awareness and participation on the part of users, recipients and consumers of AI systems. Otherwise, as Eubanks (2018) warns, there is a risk that some AI systems will not only widen social inequality through automated decision-making on the provision of social services, but may also be used as tools of social control and punishment for the poorest and least well-off people in society.

Similarly, the application of ethical values should serve to prevent AI technology from being used in what the same author calls "low-rights environments", referring to the testing grounds that different organisations may use first for the poor but eventually for all. In this sense, the contestability of the outcomes of AI or algorithmic decision systems appears as a key requirement, especially when they are used to make decisions with

a high impact on people as would be the case in areas such as health, social services, judiciary and financial services. It should be underlined that this need is recognised, to some extent, by the EU General Data Protection Regulation (GDPR), which states that a person who is 'subject to a decision based solely on automated processing' has 'the right to obtain human intervention by the controller, to express his or her point of view and to challenge the decision' (Article 22 (3)). But as noted (Henin and Le Métayer, 2021), there may be many barriers to the exercise of this right, especially because of the practical difficulty in understanding the rationale for a decision based on AI systems.

In this context, there are some proposals for assessing the potential and actual social impact of AI systems for various institutions (private companies, government agencies and civil rights advocacy organisations) to analyse and act in situations where measured impacts do not capture potential real-life harms. For example, at the public level, we know that AI-based social assessment technologies for social service delivery categorise present and future human behaviour on scales such as lawful/fraudulent, deserving/undeserving, needy/unneedy or acceptable/unacceptable recipients. Delegating decisions of this kind through such value judgements to machines or AI systems raises ethical and social issues, and leads to important questions of responsibility, accountability, transparency and also the quality of societal decision-making about the allocation of scarce resources. It should therefore come as no surprise that public opinion and discourse in these areas is highly emotive and emphatic, because fundamental societal values are affected and at stake, and decisions about the provision or rejection of public social services can and do have far-reaching consequences for the people concerned.

Undoubtedly, when the use and implementation of AI raises these ethical and societal issues an approach that involves multidisciplinary teams of researchers, practitioners, policy makers and citizens is needed to maximise equity and transparency (Lepri et al., 2018). Such a participatory approach involving many relevant stakeholders, including a multidisciplinary framework for comparing empirical cases, is likely to become increasingly important. Since the aim is to create a better, i.e. more accountable AI, one would expect civil society to have much more to offer than simply being the "moral voice" of society in research and innovation (Ahrweiler et al., 2019).

Thus, a co-construction approach can be used so that there is a real, domain-specific interactive process (e.g. health, education, insurance or surveillance) between the developers and employers of AI systems and the communities that are subject to them. The aim of such approaches is to co-construct a generic dialogue protocol that can be practised within a specific domain of social relevance and can be useful in the design,

use and implementation of an AI system. Without the adoption of such approaches, there is a very real possibility of power imbalance between those developing and deploying AI systems and the communities that are subject to them, and this situation can be further amplified when historically marginalised and under-represented groups are not involved. Common characteristics of such projects, such as Assembling Accountability[15], Artificial Intelligence for Assessment[16] or the European Network of Living Labs[17], are based on the need for sources of legitimacy, participation and forum, catalytic events, timeframe, public access and consultation, among others.

[15] Moss, E., Watkins, E. A., Singh, R., Elish, M. C., & Metcalf, J. (2021). Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. Data & Society. https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf (accessed 27 August 2021).

[16] AI FORA (2021). Artificial Intelligence for Assessment, https://www.ai-fora.de/ (accessed 27 August 2021).

[17] ENoLL (2021). European Network of Living Labs, https://unalab.eu/en/project-partners/enoll (accessed 27 August 2021).

1.10. A proposal for a regulatory framework of AI in the EU

In the face of the rapid technological development of AI, in April 2021, the European Commission circulated the proposed regulatory framework *Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence* (Artificial Intelligence Act) and *Amending Certain Union Legislative Acts* (European Commission, 2021), with the main objective of harmonising the rules governing AI technology in the EU in a way that addresses ethical and human rights concerns. The proposal details the following four specific objectives around the regulatory framework:

1. **Ensure that AI systems** placed on the EU market and used **are safe and respect existing legislation** on fundamental rights and EU values.

2. **Ensure legal certainty** to facilitate investment and innovation in AI.

3. **Improve governance and effective implementation of existing legislation** on fundamental rights and security requirements applicable to AI systems.

4. **Facilitate the development of a single market** for lawful, safe and reliable AI applications and avoid market fragmentation.

To achieve these objectives, the regulation follows a risk-based approach, differentiating between uses of AI that create (i) an unacceptable risk, (ii) a high risk and (iii) a low or minimal risk. In this sense, the list of prohibited practices comprises all those AI systems whose risk is considered unacceptable as infringing EU values. Taking into account the risks considered unacceptable:

1 The following artificial intelligence practices, including the placing on the market, putting into service or use of an artificial intelligence system, shall be prohibited:

a one which deploys subliminal techniques beyond a person's awareness in order to materially distort a person's behaviour in a way that causes or is likely to cause physical or psychological harm;

b one which exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, with the aim of materially distorting the behaviour of a person belonging to this group in a way that causes or is likely to cause physical or psychological harm to that person or to another person;

c one which by or on behalf of public authorities for the assessment or classification of the reliability of natural persons over a certain period of time on the basis of their social or known or expected behaviour taking into account personal or personality characteristics, with the social scoring leading to one or both of the following:

(i) the prejudicial or unfavourable processing of particular individuals or entire groups of individuals in social contexts that are unrelated to the contexts in which the data were originally generated or collected;

(ii) prejudicial or unfavourable treatment of certain individuals or entire groups that is unjustified or disproportionate to their social behaviour or seriousness;

---

d one which uses "real-time" remote biometric identification systems in publicly accessible areas for law enforcement purposes, unless and to the extent that such use is strictly necessary for one of the following purposes:

(i) the targeted search for specific potential victims of crime, including missing children;

(ii) the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or of a terrorist attack;

(iii) the detection, tracing, identification or prosecution of a perpetrator or suspect of an offence referred to in Article 2(2) of Council Framework Decision 2002/584/JHA62 and punishable in the Member State concerned by a custodial sentence or detention order for a maximum period of at least three years, as determined by the law of that Member State.

2 The use of "real-time" remote biometric identification systems in publicly accessible areas for law enforcement purposes for any of the purposes referred to in point (d) shall take into account the following elements:

a the nature of the situation giving rise to the possible use, in particular the severity, likelihood and extent of the damage caused in the absence of the use of the system;

b the consequences of the use of the system for the rights and freedoms of all persons concerned, in particular the severity, likelihood and extent of these consequences.

3 The use of "real-time" remote biometric identification systems in publicly accessible areas is subject to prior authorisation by a judicial authority or by an independent administrative authority of the Member State of use, issued upon reasoned request and in accordance with the detailed rules of national law. However, in a duly justified emergency situation, use of the system can be initiated without authorisation and authorisation can only be requested during or after use. The competent judicial or administrative authority shall only grant authorisation when it is satisfied, based on objective evidence or clear indications presented to it, that the use of the real-time remote biometric identification system is necessary and proportionate to the achievement. In deciding on the application, the competent judicial or administrative authority shall take into account the elements set out in the prohibition.

4 A Member State may decide to provide for the possibility to authorise in whole or in part the use of "real-time" remote biometric identification systems in publicly accessible areas for law enforcement purposes within the limits and under the conditions listed. This Member State shall lay down in its national legislation the detailed rules necessary for the application, issuing and exercise of the authorisations referred to, as well as the supervision. These rules also specify in respect of which of the listed purposes, including in respect of which of the mentioned offences the competent authorities may be authorised to use these systems for law enforcement purposes. In addition, the

use of "real-time" remote biometric identification systems in publicly accessible areas for the purpose of law enforcement for any of the purposes mentioned in point (d) must comply with necessary and proportionate safeguards and conditions in relation to the use, in particular with regard to temporal, geographical and personal limitations.

Although this proposal for an AI regulatory framework represents a step forward from the current situation of no regulation, the draft arguably fails to deliver on the main promise made by the European Commission itself in its *White Paper on AI* when it mentions that "the main risks related to the use of AI relate to the application of rules designed to protect fundamental rights (including personal data and the protection of privacy and non-discrimination), as well as security and liability issues" (European Commission, 2020: 10). However, the proposal only presents a risk-based approach that would focus on operational risks and those related to external factors and, as ECNL (2021) points out, only plans to limit oversight and safeguards for AI systems considered "high risk". In this sense, the proposed regulation not only creates a loophole for all other AI systems not considered "high risk", but also creates an uneven playing field as it unduly places the burden of proof on anyone subject to an AI system, as opposed to those developing or implementing it, and incentivises the classification of AI systems as low risk to avoid further regulation.

Certainly the European Commission's draft AI regulation proposes to ban some systems deemed unacceptable. This includes a wide range of artificial intelligences that could manipulate our behaviour or exploit our mental vulnerabilities. Likewise, AI-based social scoring and indiscriminate surveillance systems will not be allowed. As we have seen, these versions are currently being used in China's public spaces, where citizens are tracked and evaluated to produce a trustworthiness "score" that determines whether they can access different services such as transport or public employment. There is also a cautious approach to a number of AI applications identified as high risk. Among these technologies are large-scale facial recognition systems, which are considered easy to implement with existing surveillance cameras and which, thanks to this regulation, will require a special permit for implementation.

Moreover, many systems that are known to contain biases are also classified as high risk in the regulation proposal. For example, AI that assesses students and determines their access to education will be strictly regulated, which was not the case when an algorithm unfairly determined the grades of UK students in 2020. The same caution will apply to AI systems for recruitment purposes, such as algorithms that screen applications or assess candidates, as well as financial systems that determine credit ratings. In the same vein, systems that assess citizens' eligibility for welfare or legal aid

will require organisations to carry out detailed assessments to ensure they meet a new set of EU requirements

crediticias. En esta misma dirección, los sistemas que evalúan la elegibilidad de los ciudadanos para recibir asistencia social o asistencia judicial requerirán que las organizaciones realicen evaluaciones detalladas para asegurarse de que cumplen un nuevo conjunto de requisitos de la UE.

But other AI systems that are seen as incompatible with human rights, such as emotion recognition technology, biometric categorisation for the purpose of predicting ethnicity, gender, sexual or political orientation, and risk assessments for criminal justice and asylum, are not included in the current list of the European Commission's draft AI regulation. According to Europe's top data protection authorities (EDPB and EDPS, 2021), such AI systems for the automated recognition of human characteristics in public spaces and certain other uses of AI that may lead to unfair discrimination should also be prohibited as reflected in their joint communication of 21 June 2021[18]. The European Data Protection Authority (EDPB and EDPS, 2021) has also stated that AI systems for the automated recognition of human characteristics in public spaces and certain other uses of AI that may lead to unfair discrimination should also be prohibited.

In summary, although the proposed EU AI regulatory framework constitutes a global advance in horizontal regulation around AI systems, the proposed law not only risks being extraordinarily broad in scope, but could restrict legitimate national attempts to manage the social impacts of the uses of AI systems in the name of free trade (Veale and Borgesius, 2021). In this regard, the same authors point out that such a proposal gives a disproportionate role to bodies such as the European Committee for Standardisation (CEN) and the European Committee for Electrotechnical Standardisation (CENELEC), and given that AI providers will de facto follow these standards when conducting compliance assessments, other external stakeholders such as the wider civil society, academics and affected communities should be included.

Otherwise, the proposal does not sufficiently address the serious power imbalance that exists between those who develop and deploy AI systems and the communities that are subject to them, and this imbalance is particularly acute for historically marginalised and underrepresented groups. Thus, NGOs such as ECNL[19] and AlgorithmWatch[20] highlight the need for a comprehensive, inclusive and transparent human rights impact assessment in the proposed regulation of AI as a starting point for all subsequent

[18]European Data Protection Board (2021). EDPB & EDPS call for ban on use of AI for automated recognition of human features in publicly accessible spaces, and some other uses of AI that can lead to unfair discrimination, https://edps.europa.eu/system/files/2021-06/EDPB-EDPS-2021-13-Artificial-Intelligence_EN.pdf (accessed 27 August 2021).

regulatory action on any AI system. They also note that an effective right to redress for affected groups must be added to the proposed EU AI regulation, with meaningful support and adequate resources provided to stakeholders to enable them to fully exercise this right.

[19] ECNL (2021). Position statement on the EU AI Act. ECNL - European Center for Not-For-Profit Law, https://ecnl.org/news/ecnl-position-statement-eu-ai-act (accessed 27 August 2021).

[20] Reinhold, F., & Müller, A. (2021). AlgorithmWatch's response to the European Commission's proposed regulation on Artificial Intelligence – A major step with major gaps, https://algorithmwatch.org/en/response-to-eu-ai-regulation-proposal-2021/ (accessed 27 August 2021).

1.11. By way of conclusion to the first part

Throughout this first part we have reviewed the state of the art around the development of AI, as well as the main ethical and social impact considerations. AI is a technological breakthrough that affects our daily lives, whether we are aware of it or not, and the paper attempts to reflect both its implementation and its ethical and social repercussions. On a personal level, professionally and in our social interactions, with companies and public administrations, AI has become indispensable, quietly transforming the society we live in. But the knowledge we have about the specificities of AI, and what this means for the values and principles that have underpinned our society so far, remains limited. We also do not know what the scope of AI will be on various social and economic issues, for example in relation to the future of work or the distribution of wealth. What we do know is that, in geopolitical terms, specific regions of the world and specific countries view the development of AI differently, with very marked differences in terms of financial investment and also in terms of its use, including in the military sphere.

Understanding the context in which we find ourselves helps us to understand the great expectations that have been generated by this technology, and also some of the problems associated with it. We are at a very early stage of AI, and there is still a long way to go before we reach what we can consider true artificial intelligence, if we reach it at all. Powerful and accurate calculation and prediction tools based on Big Data have been developed. In this sense, if AI-based decision making fulfils its promise, we could promote economic justice through AI that enables a better distribution of resources and opportunities and more broadly from the public sphere. It is also to be expected that AI could produce substantial benefits for consumers, including mitigating some of the pervasive biases that exist in decision-making. However, AI's ability to make such transformations also risks serious harm if its use is not responsible and people-centred. Multiple examples demonstrate that AI contributes to excessive surveillance, biased risk assessments and discriminatory outcomes in a variety of high-risk economic spheres, including employment, credit, health and housing. Throughout this first part of the report we have reviewed some of these examples at the international level, but this does not mean that they do not exist at a more local or regional level, as shown in the report *Intel·ligència Artificial - Decisions Automatitzades a Catalunya* (Autoritat Catalana de Protecció de Dades, 2020).

It is therefore important to note that not all applications offer optimal solutions, and some on the market not only romanticise their capabilities, but also raise growing concerns about their infringement of fundamental rights. This is problematic, as we are beginning to define both the future development and the narrative of AI itself, a combination where risks and false capabilities dominate could affect the way we see and understand technology based on AI systems. We have the ability to define how we would like this technology to work and to anticipate the positive and negative impact that its development may have. For this reason, we must adopt policies and regulatory measures that leave no one behind. Encourage the development of tools that take care to include the whole of society. Especially the most vulnerable groups that are currently the most affected by the use of this technology, either because of over-representation as part of biases or because they do not have sufficient power or knowledge to exercise control over the data that these systems obtain from them.

As stated on the first page of the European Commission's *White Paper on AI*, the European Union has set out a common roadmap for cooperation on AI in the face of global competition to address the opportunities and challenges of AI, with the intention of defining its own path, "based on European values to promote the development and deployment of AI" (European Commission, 2020: 1). It is from these coordinates that the European Commission has produced the first ever legal framework on AI, which addresses the risks of AI and positions Europe to take a leading role globally. But despite this progress through an identification of risks and unacceptable uses of certain AI systems, the proposed regulation has important gaps and omissions. For example, the regulation does not contain a focus on the social impact of AI systems including, among others, a general requirement to inform persons undergoing algorithmic assessments, which means that compliance assessments may end up as internal processes and not documents that the public or a regulator can review. On the other hand, the proposed regulation does not strictly prohibit real-time remote biometric identification systems, and leaves the door open to their use by law enforcement when authorised by law. This loophole is of particular concern to civil liberties advocates. Thus, for example, the *Reclaim Your Face*[21] initiative launched in 2020 by civil society organisations across Europe are gaining widespread social recognition for banning facial recognition in public space.

Of course, the fact that AI systems are increasingly being used in all areas of our daily lives also implies a growing interest and concern, especially when AI is used at ethical and legal boundaries, for example to monitor and supervise protests or to make predictions about our behaviour, among others. Thus, without clear safeguards, some

[21] VVAA (2020). Reclaim Your Face, https://reclaimyourface.eu/ (accessed 27 August 2021).

AI systems may foster an imbalance of power between those who develop and use AI and those who are subject to the technology. This is why the proposed regulation of AI can be seen as a step forward, but at the same time, the possible contexts of action must also be taken into account. In other words, the interpretation of different technical applications may not only differ according to a risk assessment of AI, but is also subject to the judicial system itself and the different values and norms of each context or country. In this sense, the development and deployment of AI could vary depending on whether or not a policy of maximum democracy is applied to protect and safeguard both civil rights and the plurality and diversity of its citizens. It should therefore be anticipated that oversight and enforcement of AI regulation will be complex due to the inherent EU division of responsibilities between regulators in Brussels and Member States, and also because national supervisory authorities are expected to take the lead in the so-called 'market surveillance' of AI systems.

We know that AI has enormous potential to help us as demonstrated in the Covid-19 crisis. And it is not the only far-reaching example. The achievement of each of the Sustainable Development Goals could depend on the use of technology and especially AI as demonstrated by multiple UN AI for Good initiatives. But as we have seen, AI not only presents solutions and opportunities with its growth, it also presents implications and concerns, and we must be aware of the importance of its limitations and strict prerequisites, otherwise AI can also have a very harmful impact as the authors of the Barcelona Declaration themselves underline. Thus, Steels and Lopez de Mantaras (2018) call for a principle of prudence as, on the one hand, the application of knowledge-based AI requires the availability of sufficient human expertise and resources for detailed analysis and, on the other hand, data-driven AI requires sufficient high-quality data and a careful choice of algorithms and parameters for each case. Thus, exercising the necessary prudence in the application of AI by those responsible for its development appears to be fundamental. But the implications of AI are not just a matter of internal development and validation, they also concern everything surrounding its external development as an industry, and its relationship with political power.

From this perpective, we are increasingly aware that corporations, consultancies and states, often acting in concert, are drawing and enacting an AI future based on specific profit and security interests (Amoore, 2013). From the minerals extracted from the earth to make it work, to the labour extracted from low-wage information workers, to the data extracted from every action and expression of each of us, AI is increasingly seen as an extractive technology (Crawford, 2021). For this and other reasons, we need an ethics of doubt regarding current trends towards algorithmic governance (Amoore, 2020). Therefore, at this spring stage of AI, it is important to ask what AI we want and for

what purpose, and to ask: can AI solve any problem? The answers to these questions depend, in large part, on ethical considerations we make ourselves, and also on the societal impact we are prepared to take on as AI becomes ubiquitous.

Today, the technological progress of AI is rapid and genuine in many fields, from image perception to the automation of criteria, but none of them is free of errors and inaccuracies, especially in the prediction of behaviour of a social nature. But regardless of whether we reach high degrees of accuracy in many disciplines with the use of AI, the central question is not whether AI can do one thing or another, the central question is whether it should do and how. So delving into the motives and intentions of AI development is critical, both to establish whether or not AI needs to be used, and in what contexts and circumstances. And, of course, if we use AI systems we also need to anticipate that this implies (increasing) scrutiny of whether the benefits of doing so extend beyond private interests, i.e. to individuals, groups of people or society at large. Therefore, various questions (positive and negative) towards AI need to be studied and analysed, such as: can AI give me better options, can AI save me a lot of time, can AI harm me and others, could AI eventually replace or restric me, and if so, how?

It is through this understanding that we can know what is good or bad, what can and should be done and also how to do it. We know that this differs both from philosophical or scientific knowledge (*episteme*) and from technical knowledge or know-how (*téchne*). First, because it focuses on practice, which means that it is not only about what is true, but also about what it would be good to do in certain circumstances. Second, because it is as concerned with assessing and prescribing objectives as it is with selecting means. Without such an understanding and coordinates for action, it would not be surprising that there is a growing distrust and an equally strong demand to set highly regulatory limits to AI deployments, both at national and European level.

Like any technology, AI systems often distribute benefits and harms unequally, and also aggravate or perpetuate pre-existing unjust social conditions. In this sense, public administration should be a clear benchmark for the development of ethical and people-centred AI. It is clear that AI can help a great deal in managing and analysing administrative *Big Data*, offering advantages such as a faster and more transparent service, while anomalies or fraud can be detected. But, above all, it must be ensured that AI is used without causing any disparities with regard to social rights and social cohesion. An allocation of welfare and eligibility assessment resources cannot be based on risk prediction AI as this has a potentially serious impact on the fundamental right to social security and social assistance.

Finally, it should be stressed that in both the public and private domains, AI systems can only be as good as the data used to develop them. High quality data is therefore essential for high quality AI systems. Unfortunately, the current reality of AI is still far from widespread use of high-quality data. While it is impossible to have error-free data, it is possible to know the various sources of error in all data collections, and users of AI-related technologies need to know where the data comes from and its potential shortcomings. This is a critical aspect as AI systems based on incomplete or biased data can lead to inaccurate results that infringe on people's fundamental rights, including discrimination rights. Therefore, being transparent about what data is used in AI systems is a first step that can help avoid potential rights violations, and this is especially important today with *Big Data*, where very often the volume of data is valued over the quality of the data.

At this point, we must continue to work and cooperate to demand ethical technological models that, in turn, allow us to advance and achieve innovation in the field of AI. Listening to the different proposals or perspectives given to us by experts through consequences or opportunities can help us define the direction to take to achieve this goal. Therefore, in the following section of this report we take a look through the opinions and reflections of different experts and/or stakeholders on the risks, opportunities and open debates for the adoption of ethical AI from a social and legal perspective and also with their vision or look into the future.

PART 2

2.1. Collecting and analysing qualitative information

"SCIENCE AND EVERYDAY LIFE CANNOT AND SHOULD NOT BE SEPARATED"
ROSALIND FRANKLIN

Between February and April 2021 we conducted a total of 23 interviews[22] with experts and/or people interested in the development and impact of artificial intelligence from different fields, ranging from academia to industry and including public administration and citizens. The main objective of these interviews was to find out their opinions and reflections on the current development and implementation of AI systems, taking into account different ethical, social and legal aspects. With this qualitative work, we also wanted to capture the prospective or future vision of the main challenges and opportunities that arise in relation to ethical and social AI. In this sense, the interviews focused on three areas of interest: (1) the ethical and social domain, (2) the legal domain, and (3) looking into the future.

The selection of interviewees was intentional and based on their potential for richness of information. That is, we intentionally sought out a range of people considered to be experts who could offer rich and accurate information on the topics under study. While their expertise is based on their experience in a specific area, the interview scripts were shaped so that the interviewees would address a broad audience, and consider important ethical, social, legal and political issues of the day around AI that go well beyond their professional interest or field of work. Following the work of Baert and Morgan (2018), they can be considered experts and informants according to the same terminology we have used in this paper.

The design of the interviews was semi-structured with the aim of facilitating a two-way communication and allowing the interviewees (informants) the freedom to express their points of view in their own terms while, at the same time, delving deeper into different topics according to their knowledge, experience and interest.

Semi-structured interviews are frequently used in the social sciences to obtain reliable and comparable qualitative data through a general

[22] For a compilation of videos and audios of the interviews conducted, please consult this YouTube link (https://youtube.com/playlist?list=PLEqIWgQEwkLsf-QyTSxH3m8N1olztwwOt)

structure that allows for the collection of key information and discovery across the three thematic trajectories mentioned above. All interviews were conducted in a videoconference (virtual) format using the support of the Zoom platform to achieve this (with only one exception where the interview was face-to-face and digitally recorded). Semi-structured virtual interviews allow the interview to closely resemble the natural back-and-forth of face-to-face communication since, among other things, verbal and non-verbal cues can be perceived as the communication between the interviewee and the interviewer is via audio and video. However, the fact that the communication was via the Internet means that there is a mediation of the information, hence this type of information gathering is known as Internet-mediated research (Hewson, 2010). Despite this mediation, we can consider the medium used as an advantage considering the restrictive factors of the Covid-19 pandemic and the fact that it has also made it easier for us to contact people who might otherwise be unreachable.

The semi-structured virtual interview script used a total of 12 open questions (6 from the ethical and social domain, 4 from the legal domain and 2 looking into the future of AI) and 6 closed questions (2 from the ethical and social domain and 4 from the legal domain). The latter were questions in which a Likert scale of 1-5 (1 being strongly disagree and 5 strongly agree) was used for coding the response. In order to obtain a broad understanding of the different areas of interest, the interview questions were framed as either stimulating or challenging. At the beginning of all semi-structured interviews, two situational introductory questions were asked to find out what discipline or field of AI they work in or have an interest in, as well as their view of AI through a short word or expression. Both (more informal) questions were used to develop the relevant and meaningful semi-structured questions and also to identify possible aspects to understand about the topics in question. Annex 1 contains the full script of the semi-structured virtual interview.

Although the number of qualitative interviews needed to complete this second part of the work was not specified at the outset, a total of 23 interviews were conducted after reaching data saturation taking into account the three substantive themes. Therefore, data saturation was used as the fundamental criterion for qualitative sampling. As is known in qualitative research, further data collection becomes unnecessary when such saturation is reached in terms of identifying new themes. It is generally recognised that the minimum number of interviews should be between twenty and thirty for a qualitative interview-based study to be published (Bryman, 2012: 425).

All interviews were conducted by two people after conducting two pilot sessions. The interviews proceeded as planned and the duration of each interview was approximately

50 minutes, i.e. less than an hour, which is a reasonable time for this type of interview in order to minimise fatigue for both the interviewee and the interviewer. The data collected were audio and video, although the analysis and automatic transcription of the information collected focused only on the audio files using NVIVO v.1.5 software.

Thematic analysis, commonly used for the identification and interpretation of patterns or themes within qualitative data, was used to examine the information collected in the open-ended questions. This analysis has been deployed sequentially: first with the reading and familiarisation of the content of the data collected, and second with the generation of initial themes and identification of potential themes of respondents' positions on the questions posed. From a methodological point of view, the thematic analysis has been approached from an external (respondents' information) and internal (interviewers' information) perspective, which involves generating and reviewing themes, as well as checking whether the themes generate a convincing story from the questions posed. The results obtained from the closed questions have been analysed through bar charts and are available in Annex 2. The result is the analytical narrative and contextualisation of the information from the rest of the paper, in which we have established positions on the open questions posed in the ethical and social, legal and AI futures domains of the interviewees.

As noted above, the interviews were conducted between February and April 2021 and therefore in the context of a global health crisis. The pandemic will most likely become a turning point, as it has greatly accelerated the use of AI systems and big data, in many cases proving effective in monitoring, detecting and identifying biochemical, molecular and cellular factors associated with Covid-19. But we know that there are many other factors that may have an impact on people's opinions, including institutional aspects such as the proposal for the regulation of AI systems by the European Commission in April 2021 or the inauguration and decisions of a new president (Joe Biden) in the most advanced country in AI systems. In that sense, it is worth remembering that this paper covers written responses from experts explaining their personal views and reflections at a specific point in time on the ways in which individuals, groups, organisations, countries and world regions develop and adapt AI systems to multiple challenges and opportunities.

The following section shows a selection of the answers obtained and considered to be the most complete and shared by the experts who participated in this second part of the work. Some responses have been slightly edited for style and readability, and following many official ethical guidelines, we have anonymised personal identities (names and surnames of interviewees) using a unique numeric identifier for position

analysis. However, in order to acknowledge their valuable and invaluable contribution to this work, the list of interviewees is included in Annex 3.

2.2. Ethical and social domain

2.2.1. Ethical considerations of AI: restriction, sub-objective or main objective?

We have seen that the ethical implications and moral issues arising from the development and implementation of artificial intelligence technologies can take different forms, although the aim remains to articulate general values on which we can agree and which function as practical guidelines. In order to ascertain and contrast the positioning on ethical considerations of AI we asked the following question to the interviewees: Looking to the present but also to the future, **do you think ethical AI should be seen as a constraint on AI actions, as a sub-goal or as the main goal?**

Following the question, we have grouped the answers according to whether ethical considerations are seen as the main objective, whether they are seen as a configurative part of AI, or whether they are not seen as a constraint or even as a main objective.

MAIN OBJECTIVE

[Interviewee #1] "I think in all these AI technologies up front the goal is to optimise but I would say that ethics is a primary goal."

[Interviewee #13] "I think it is the main objective and I think for any software development or engineering in general, no, we are talking about engineering, software or telecommunications, but if we talk about any engineering in general, the purpose of the purpose of why you are doing an engineering work is for something. So the purpose has to be within the context. It's not something separate, it's within the design itself."

[Interviewee #14] "I think the answer depends on the application. That is, only in applications that affect people. I think it should be an objective, I don't

know if it is the main one, but a very important one."

[Interviewee #17] "Undoubtedly ethics and therefore values that prioritise the goals we want to achieve as a society have to be at the very beginning in the design and therefore it is a fundamental part of artificial intelligence and I have no doubt that as the main objective."

[Interviewee #21] "Look, I think it is a fact, a main objective. I think it is a responsibility that people who are involved in this area have to bear in mind. I mean, we do things to generate benefit for society, to generate satisfaction, to generate future wellbeing."

CONFIGURATIVE PART

[Interviewee #2] "I think it has to be a configurative part of artificial intelligence. It can't just be restrictive. If you want artificial intelligence you have to understand the ethics of artificial intelligence because from the beginning this discipline has taken ethics into account."

[Interviewee #3] "I believe that the ethical framework should be the starting point, but we have to define what is ethical, what are the basic principles, which are common to all actors. I always say that our framework first is what we are doing in Europe. The big challenge is to first define what these ethical principles are and how you transform them into operational aspects that you can force within a regulatory framework. Another thing is that we don't like the ethics that this discipline has had and we want to change the ethics."

[Interviewee #4] "I don't see ethical artificial intelligence as something restrictive. That is, I don't think it has to be seen as restrictive, because artificial intelligence or the advancement of artificial intelligence should also be about the direction or purpose that we have with artificial intelligence. In fact, I have always said that I really like Professor Stuart Russell's vision when he talks about beneficial AI. I am going to tell him that in the end the goal of artificial intelligence should be to have a positive impact on the human being, even without knowing or even without taking it for granted, that the human being understands very well what is the main goal that the algorithm is looking for, because many times we think that this goal is very well defined and in reality it is not. On that basis, for me, ethics is not a restriction or an objective, but

rather a characteristic that artificial intelligence should have."

[Interviewee #22] "I am not convinced that ethics has to be central to the whole development of artificial intelligence. However, we must keep ethical considerations in mind when deciding even what aspects we work on and research. In general, we should ask ourselves questions about whether what we are doing is worth doing. And if it's not appropriate to do it, to be able to make the decision not to do it."

NEITHER RESTRICTION NOR MAIN TARGET

[Interviewee #6] "Obviously not as a constraint, because that is not what some people think, that ethics brings limits and problems, not as a main objective, which would even be considered a bit exaggerated as a super objective. Nevertheless, we can make a law in the sense that we humans find the ethical perspective to be primary. Although I would say that it always has to be principal, what happens is that if we apply it to artificial intelligence I think we have to recognise that its qualitative construction is also principal, we would say. I would say that it has to walk on two legs, the leg of the efficiency of the product and the leg of its adaptation to human interests."

[Interviewee #8] "Look, I think that ethics has to be part of the relationship that is established between humans and artificial intelligence. In other words, I see it more in how this artificial intelligence is designed to allow a relationship between humans and artificial intelligence. [...] I don't think it has to be lived as a constraint. I don't think it has to be seen as a limitation but surely as a style of doing artificial intelligences."

[Interviewee #11] "I don't see it as a main objective and it should not be seen as a constraint. If it has to be, let's leave it as a sub-objective."

[Interviewee #15] "I deny it. I mean, there is no such thing as ethical artificial intelligence. What there is is ethical uses of a technology and that is for any technology. So it is absurd to say that artificial intelligence is ethical because it has no agency. The moment we could create autonomous, independent, responsible agents with rights, we could start"."

[Interviewee #16] "I think it will work, but what we cannot do is ask of artificial intelligence what we have never asked of people throughout history. In other

words, what we cannot do is stop artificial intelligence because it is not politically correct. Yes, but without drama and without going too far. The answer is yes."

2.2.2. AI as a factor in human debilitation

There is a common understanding that AI is to serve to help humanity solve problems and facilitate multiple work processes, whereby through AI we have a tool that emulates the "cognitive" capabilities of natural intelligence. In fact, AI, together with the rapid development of cyber technology in recent years, is so commonplace in our daily lives that we are very accustomed to it, especially for navigating physically and virtually with information-seeking equipment. It is for that reason that we asked the following question to the respondents: in your opinion, **do you think that AI will weaken or discourage some important human habits, skills or virtues that are fundamental to human excellence (moral, political or intellectual)?**

Through the information obtained, we have grouped the responses based on three approaches of the interviewees, which would be the positive, negative and neutral or ambivalent position (expectation).

POSITIVE POSITION

[Interviewee #2] "Let's see, this is debatable. Artificial intelligence was born as a tool to help humans perform certain functions. And so it is, in general, for all digital technologies. [...] That this has resulted in artificial intelligence that in some way aims to supplant human capabilities is another problem. Therefore, the development for the future is open. That is to say, it is open in the sense that official intelligence can indeed help develop a whole set of human capacities or it can cancel out a large part of these capacities. But it is open, and I suppose that ethical artificial intelligence has to respond to this openness. [...] I think that now artificial intelligence designed by a set of institutions, not only academia but also, as we say, the quadruple helix, can reorient a new stage of artificial intelligence connected to the development of these human capacities."

[Interviewee #8] "I think it is a technology that is not inherently a technology to undermine the human and it is a technology inherently created to

empower and to collaborate and to enhance the capabilities of the human."

[Interviewee #6] "I have the conviction that it will change us, but I would say that it is still too early to know in what sense. And I do intuitively [...] believe that it will change us and that it will replace deficiencies that we have and improve our possibilities. And obviously if the ethical perspective ends up being real, then the result will be positive."

[Interviewee #11] "No, I don't think so. I think it's going to be a complement, just as it has happened with other technologies. [...] And you have to say that a machine or an algorithm, or a technology is giving you support to make, to advance, let's say, in a more efficient way. I find it discouraging, I find it, the truth is."

[Interviewee #13] "That we humans can do such beautiful things as this with software development, which is my profession, is because we are moving towards excellence, i.e. it is a means and not an end."

[Interviewee #16] "Absolutely not, I believe that we will learn how to manage artificial intelligence and it will take us, as all the technological tools, to a stage of augmented intelligence."

[Interviewee #20] "I don't see this happening, on the contrary, it will strengthen us. [...] Perhaps technology will make our jobs change and we will dedicate ourselves to other things, and in the long term this may condition the skills we develop or not. But I believe that this trend has to move towards a more evolved society with higher and higher ethical criteria."

[Interviewee #21] "Of course, the question is very much conditioned by what we mean by important for the human species. I believe that what is essential cannot be replaced. [...] Questions such as those I mentioned a moment ago, creativity, innovation in real terms, meta-knowledge and the relationship between the concept of the self and the concept of society. All this is human. And this for the moment, at least, cannot be substituted."

[Interviewee #23] "I don't think it should weaken us; on the contrary, it should empower us. We are starting from a higher point and I don't see that AI is

different from any other technology. There may be changes in some of the capabilities we have, we may not need them if they are given."

NEGATIVE POSITION

[Interviewee #1] "I don't know if it only weakens us, but what is clear is that it changes us. That is to say, technology shapes us, going back to one of my novels. The leitmotif is the relationships that we build, which in turn shape us. [...] The increasingly close relationships we have with technology are shaping us, and it conditions the future and future generations even more. So, of course, some skills are decreasing, such as spelling, calculation skills, because they are clearly decreasing. Perhaps others are on the rise, such as attending to several things at the same time."

[Interviewee #7] "Maybe in the short term maybe not, but in the long term when the technology develops further, let's say machines can become smarter. It is possible that it could exert or could have a disincentive effect."

[Interviewee #10] "I think there is a possibility. I don't dare to say whether it will or won't because it depends very much on the implementation, it depends very much on the ability we have to get it to the people who need it. [...] Now, I do think there is promise that artificial intelligence will allow us to develop certain capabilities. [...] What I do think is that we would have to regulate and incentivise innovation around artificial intelligence that would really allow us to implement this in a positive way."

[Interviewee #14] "I believe that limitation exists, but it comes from ourselves. [...] We limit ourselves by feeling helped by technology. For example, people don't know how to read a map anymore, because they use GPS all the time. But yes, we have infallibility because we learned hundreds of thousands of years ago. The same goes for other things, like memory. We trust that the contacts are going to be there, in the mobile phone, all the time. We used to remember them. Maybe not so many, but we remembered a lot of them. I think we limit ourselves because we get used to it. Not because technology is limiting. So the limitation comes from the same cognitive biases that we have when we use technology. [...] I think that's a problem because we are losing ancestral skills that were evolutionary and that can be lost very quickly. These are skills that have to be practised."

[Interviewee #15] "Yes, using media, many of the media on a daily basis and out of laziness will limit you. I always do this, I ask my students some questions and I ask you, how many telephone numbers do you know by heart? How many addresses do you know completely? And the million dollar question, how many varieties of apples do you know - red, green and yellow? So, of course, what we are doing is making it easier for people to do more tasks in a simpler way, but we are taking away from their capabilities and we should not believe that it is not good if a human cannot multiply, divide, take square roots."

[Interviewee #16] "What is more worrying is whether all this [...] dehumanises us, this is the key here. And yes, there is a danger and perhaps a disincentive that affects our direct social relations, of people with other people. There are plenty of examples of an absolutely wrong use, in my opinion, of technology. Indeed, I am very concerned about this misuse of these technologies in the sense that not only can they cause us to lose certain skills per se, but they dehumanise us and treat us almost like machines."

NEUTRAL POSITION OR AMBIVALENT

[Interviewee #3] "Well, I think that any technological development impacts us as a species, as a collective in the relationship between this same collective, right? So I'm going to tell you that artificial intelligence is going to have an impact, but it's going to have an impact like it had when electricity arrived, when we started to communicate, when steam and trains arrived, which changed our experience in relation to space. [...] And now the question is what impact we want it to have."

[Interviewee #9] "I would like to point out that artificial intelligence as such does not do anything, so it is the people behind the programmed artificial intelligence that could do all the goals that artificial intelligence sets."

[Interviewee #12] "It depends on the framework in which we put it. It can enhance certain human actions and it can be a disincentive to them."

[Interviewee #17] "I think that like everything else in this ethical reflection, my contribution is not what we do but how we do it. In what direction we give it and why. Until then I think it has its positive and negative drift, the pros

and cons of the technology. And that's why we have to do this analysis before and during the development."

[Interviewee #18] "I don't know if it will weaken them or strengthen them. That depends on how it is carried out, and it can change them and technology has always done that. That is, it has changed the way we relate to each other and to the world in general. So that's an inescapable thing."

2.2.3. AI as a factor of human empowerment

History tells us that human beings are always looking for ways to be more efficient, to go faster, to be more effective and to seek more convenient and even more personalised solutions. In that sense, we can ask ourselves whether AI presents itself as a factor strengthening these human ambitions or whether, on the contrary, people are already satisfied with a natural way of life without excessive desires to progress through AI technologies. It is for this reason that we asked the following counter-question to the respondents: **and on the contrary, do you think that AI will strengthen some important human habits, skills or virtues that are fundamental to human excellence (moral, political or intellectual)?**

The answers obtained not only allow us to group them in a similar way to the previous one (positive and neutral or ambivalent position), but also to gather new information and nuances with respect to a contrary formulation.

POSITIVE POSITION

[Interviewee #1] "These new technologies are used in very different ways by different groups. I mean that some people are experiencing the digital divide or the digital divide and there are people who use artificial intelligence or the new digital tools for their own benefit to cultivate themselves, enrich themselves and gain much more knowledge and go faster. There are many others who use it only as a tool for entertainment and it doesn't help them, on the other hand, perhaps it diminishes certain capacities, but I believe that those who use technology well have a brutal multiplying effect. [...] I would like us to move towards, I won't say a symbiosis, but a complementarity between what the machine does better because it is unquestionable that the calculation capacities, the speed and a whole series of things surpass me. There

is no problem with this, but there are others that we can excel more than the machines and so these are the ones that we have to try to cultivate and enrich. Therefore, what I like is the union or fusion of man and machine in the future to bring out the best of each of the same capabilities."

[Interviewee #2] "Yes, it is an important part, in the same way that simpler calculators helped increase the ability to do simple maths. Artificial intelligence can help to develop complex cognitive operations, i.e. there is not necessarily a contraposition or substitution but a complement to both the cognitive capabilities of humans themselves. I think this symbiotic dynamic is possible."

[Interviewee #3] "I believe that artificial intelligence, its use in certain contexts as an augmentation of knowledge, of the cognitive capacity of human beings, can have a brutal impact. For example, we are always talking about healthcare, and telemedicine operations, analysis of medical tests."

[Interviewee #5] "Absolutely. Of course it will strengthen competencies and capabilities that have to do with automation and how the human being can take advantage of an extension, right? An extension of your body, also of your mind [...] you can use this vision of an extended human being."

[Interviewee #13] "Yes, yes, I think it will strengthen habits. In fact, this is what we are working for. I think that's our mission as professionals working in this field. [...] To give a legacy that is better than the one we received back then. Yes, yes, I think it is very, very much related to the principles of who builds and who uses."

[Interviewee #14] "It should be. I think that just as we get used to it very quickly, we can get used to it in a positive way by being able to do some things that we didn't do before. I think maybe it doesn't happen much today, but I think in the future maybe the systems will help us realise things we didn't realise, for example in our own biases, and that will allow us to do that part better."

[Interviewee #16] "Absolutely yes. I think it will favour it. I think it will greatly increase the ability to make quality decisions."

NEUTRAL POSITION OR AMBIVALENT

[Interviewee #6] "I think it multiplies a lot of possibilities and can eventually replace some deficiencies. Who knows if one day there will be an artificial intelligence software mechanism that can help people who have severe memory deficts, for example. [...] And that will be a challenge because then we would have a machine helping us in a part of our intimacy that is essential because it is made of the memory of the identity of self-knowledge and recognition."

[Interviewee #9] "I think it is still too early to give an answer just as it is not too early to give an answer that we are losing. For me, it's a bit early to see the positive side because it's not so obvious. I mean, we have to spend more years living in this fast-paced world to see what it's doing for us, right? I would quickly tell you no. [...] This answer would probably be given differently by young people who are already digital natives and don't have the analogue aspect that we have, they don't have the awareness of what they are losing."

[Interviewee #11] "Well, in the end there will be. There will be a part of advancement. So yes, it will strengthen certain virtues and even new virtues. Things that I have never, that I have never reached today because of the technological development that we have until now. Now these words do not go to young people, now they are digital natives and therefore they have more facility with the use of digital technologies. [...] Technological development does not have to be an incentive or a disincentive. However, digital training does provide an incentive, or digital training does provide an incentive."

[Interviewee #12] "I really believe that everything is still to be done. It's all very open. I would say, I am expectant. [...] Now we could say that the narrative is being constructed in artificial intelligence and here the imaginaries are very important because they are going to condition a lot of how we understand it."

[Interviewee #15] "You use the tools that come from artificial intelligence to compensate for the loss of physical or cognitive abilities. What it does allow you to do is to stay autonomous longer and that's a different thing. So, you know, people are increasing their knowledge and capabilities."

[Interviewee #17] "I think it might be a supportive and helpful tool, that strengthens. But, of course, we always have to determine why, no? I do think it can help."

2.2.4. The context of ethical considerations in AI

We must recognise that when we take into account ethical considerations such as responsibility, justice or others, these can be interpreted very differently depending on the geographical and cultural context. In this sense, it is often said that good decision making, also in the field of AI, is much easier when one understands the situation or context. Understanding or awareness of the context allows the selection of a more appropriate set of characteristics of the place and the people involved or affected to reason out the best possible solution. Hence the expression "intelligent" behaviour is commonly associated with simple understanding of a situation rather than complex reasoning. To find out more about this, we asked respondents the following question: **in your opinion, do you think that the ethical perspectives of AI recipients and communities other than our own, including those who are culturally or physically far away from us, need to be considered?**

On the basis of the responses obtained, we have made a grouping between those who have shown a favourable position and those who have a rather neutral or ambivalent position.

POSITION IN FAVOUR

[Interviewee #2] "Yes, yes, radically yes. An anthropologist's answer cannot be different, precisely because we value more the cognitive diversity not only of the receivers but also of the generators themselves, i.e. the generators of this artificial intelligence also have a cultural perspective. Therefore, the more diverse the inputs from these cultural perspectives, which are cognitive, the more the diverse cultural intelligence is enriched. Yes, definitely, yes."

[Interviewee #5] "Of course you have to take into account all the considerations of the person, of the impact that it is going to generate, an impact that has to be sustainable. Hopefully it will not create tensions within the community, that does not destroy but helps to build or alleviate, to transform."

[Interviewee #7] "Certainly, when artificial intelligence-based solutions start to be used on a massive scale, the ethical implications that this may have in any field have to be considered, as well as those of the recipients or consumers. Of course, of course. We will have to make an effort."

[Interviewee #8] "What you are saying is fundamental. [...] The way you deploy the implementation of this artificial intelligence has to be fully contextualised. In my opinion, and therefore, one of the very big holes that I think there are in the area is that this exercise is not being done. [...] And I think this is the big hole where we can have an impact from the observatory. [...] And I'm not just talking about the cultural differences that may exist between Sweden and Catalonia, but imagine if a kitchen robot is deployed in a house of blind people or people with disabilities... so we're not just talking about cultural things."

[Interviewee #10] "Being permissive for me is absolutely essential [...] that is to say if we want it to work for people we cannot only take into account the cultural and cognitive mental frameworks of those who are developing it [...] obviously we have to take into account these biases these cultural sensitivities because otherwise we will be discriminating positively in a negative way."

[Interviewee #11] "Technological development does not normally stay in one country. In the end, we use many things that come from abroad and the same thing will happen with algorithms. It is very likely that an algorithm trained in China could be used in Spain. There is no law that prohibits it. So you have to take into account or make sure from an ethical point of view, especially with the issue of ethnic minorities."

[Interviewee #12] "Yes, of course. Precisely the beauty of ethics is that there is no one ethic, there are many ethics, aren't there? [...] And that's the interesting thing too, to see how we combine these different ethics."

[Interviewee #13] "Yes, the answer is yes, I am in the innovation business. We cannot innovate if we don't listen and have critical thinking and especially for two things, to listen to a diverse opinion or a diverse situation, which questions our principle, but also to reinforce it."

[Interviewee #14] "Very interesting question. [...] One of the issues I touch on is cultural

differences and they are not only to do with religion. For example, in Islam, law is a subset of ethics which makes it much easier to talk about the ethics of law. That is, if something becomes law, it does not cease to be ethical. In contrast, in the Christian world, [...] it's that if you regulate something it kind of ceases to be part of the field of ethics. And that's very strange to me, that something that was ethical ceases to be part of what is ... it loses a property. [...] In Ubuntu the most important thing is the community. I mean, nobody is what they are on their own, it's because the community exists and that's very important. I mean, we are what we are because we are in a community."

[Interviewee #17] "I think that this diversity and the representativeness of all those affected must be represented and integrated, and the social logic behind it, since it is not clear that if we are talking about artificial intelligence in an environment such as ours at European level, then it will have its own particularities."

[Interviewee #18] "Yes, definitely. Here it seems to me that all perspectives of different stakeholders need to be considered, including people who may be affected, even if they are not directly involved in the systems or even future generations."

[Interviewee #20] "I understand that yes, the addressee of everything that is developed is obviously the core of the attention and of the decisions that we have to take. From the point of view of ethical functionality or who the system has to serve or the system has to be built to serve well in an ethical way in this group."

NETRUAL POSITION OR AMBIVALENT

[Interviewee #1] "This is one thing that concerns us a lot at the research level because of course it slows us down in a way. We will make a more ethical robotic artificial intelligence but maybe the others have already sold it. [...] Obviously the cultures are very different. But I am more concerned about imports than exports."

[Interviewee #4] "I would say that the more diverse the team we are using to design the solution, the better, especially if we are thinking about a receiver in our life."

[Interviewee #6] "We need to get everyone to agree to the extent that we would have to create the broadest possible ethical consensus on artificial intelligence. Accompany this ethical consensus on artificial intelligence [...] with a radical reform of the UN Charter of Human Rights that was made seventy-two years ago with Western criteria. [...] It has to be realised, broadened and consensual, universalised."

2.2.5. The impact of AI on younger generations

As noted above, it is possible that AI could exacerbate some existing social and economic problems, for example by eliminating jobs and causing unemployment in automatable labour sectors. This type of impact may be a problem for both young people and those without a technological background. While many young people have time to gain experience and, to some extent, anticipate the impact AI will have on their lives, we do not yet know what the impact of AI will be on younger generations. It is to be expected that the market will be affected by increasing automation of jobs, and therefore, for those people who are at the beginning of their entry into the labour market, this could affect them. We are very likely to be the first generation to work side by side with AI, so acquiring relevant training and experience would be a conditio sine qua non. To find out more about the possible impacts of AI on younger generations, we asked respondents the following question: **how do you think the younger generation might be affected by the widespread use of AI systems?**

Taking into account the answers obtained, we have made a grouping that would take into consideration a positive and a rather negative position regarding the impact of AI on the younger generations. We would also like to point out that when we put this question to one of the interviewees, the answer was quite eloquent: "I think we should ask them."

POSITIVE POSITION

[Interviewee #4] "Well, I think that the younger generations are already born with the mindset, they are more digital, they are used to it [...]. I think that it is going to make them more demanding with brands or with what they expect in general, not only with brands, but with the groups, people, administrations with which they interact, because in the end their level of reference in terms of the species and the specificities of those brands that interact with them is very good, which means that they are

being very spoilt and educated by the technological giants to receive what they want and when they want."

[Interviewee #6] "Yes, it is obvious. I think we have an advantage and that is that, as the saying goes, they were born inside. [...] They are already selling the famous idea that some people are fighting against, but I think it makes a lot of sense for digital natives to move very naturally in this. [...] It's one thing for them to master the manipulation of machines, but it's another thing for them to know what's going on inside and what the consequences are. Therefore, it makes sense to ask why both young boys and girls and older people will need a culture and a continuation, a digital literacy that includes an awareness of what the algorithms that inspire all the procedures and the consequences of the application of these algorithms represent. In other words, we are inventing something that can overwhelm us negatively but that could help us very positively. [...] The question is whether we will be able to incorporate the processes of artificial intelligence into an education that does not cut back on the qualities of natural intelligence."

[Interviewee #8] "This question has many facets. I mean, first of all, I think that the generations are now interacting in an unstructured way with artificial intelligences and this may even affect their development or their personal growth, right? Point number 1. Point number 2, due to this lack of training, I believe that there is no AI culture, there are no tools or personal resources to relate to these AIs, so everyone does as God has given them to understand, right? And often without realising the consequences that this can have. And then this of course in the long run will take its toll on people. [...] And therefore I think that there will even be a transformation of the labour market where the role of the human. Thus, I think that there will even be a transformation of the labour market where the role of the human will rise in some way and therefore this has a lot to do with how we train these young people because at the time they may be less interested in driving an engine, right? And they are more interested in wanting to develop a more socially elevated role at the end of the day, eh?."

[Interviewee #9] "With this I will make the same analogy as before, in the same way that we incorporated the wheel, fire or the fact of riding a bicycle into society, yes, and therefore I suppose it was a trauma the first time a bicycle

appeared and people didn't know how to ride, but our generations, on the contrary, have the illusion of when you were a child and of knowing how to ride a bicycle. I understand that the new generations, just as they have been digitalised... so I think they are digital, they are artificial intelligence and therefore it doesn't affect their development because it's already there."

POSICIÓN NEGATIVA

[Interviewee #1] "This is the issue that concerns me the most because I think that it is at this stage that they have to learn the potential benefits and risks. And in this sense, of course, I have developed materials to teach technoethics at secondary school, high school, ESO and university level in technological careers, which are the ones who will develop technology in the future. I think it is very important that they at least acquire this critical thinking of looking at technology, seeing both the benefits and the risks, so that they are aware of the implications of what they are developing. I think this is basic, and regulation is important, but this is even more important. And if it is done well, then these generations will develop a more ethical technology, at least at the European level."

[Interviewee #9] "I think that slowly this will become more and more [...] it is like the issue of privacy, we have accepted, whether we agree with it or not, we have accepted that through mobile phones they are taking a lot of information and that this information is useful because through the apps they give us what we need. [I don't know if this will change in the future and therefore, how can they be affected? I don't even know if it will only be considered by the younger generation who feel affected by the use or just live in this world, that's my impression."

[Interviewee #10] "One of the drifts I see is, for example, when there is a whole series of dynamic decisions that are affecting their lives, they are not aware of them. Precisely because they are processes that are very well invisibilised [...] I think that [...] limits their capacity for agency."

[Interviewee #12] "I don't know how they're going to be affected and for sure, as happened with computers, eh? I'm not a digital native they call me, but there are people who already are. [...] It's obvious that to the extent that artificial intelligence has a greater weight in our lives, it's going to condition the

way in which people grow up in those environments."

[Interviewee #13] "Very much so, directly, as we have been affected, as with the development of the Internet, which in my case I am from a generation that has still lived the before and after. And the same thing is really going to happen and it's already happening, isn't it? We are all already affected by the development, which is becoming more and more accelerated."

[Interviewee #14] "Yes, I think so. And the issue is that it is not only in the future. One issue that is now and is getting stronger and stronger is privacy. [...] Because nowadays privacy affects young people a lot. All the examples of people who didn't care much about what they put on social networks when they were young and then they realise that it's all public."

[Interviewee #15] "They are already affected. Ah, they are going to be very affected, fundamentally because those who have allowed them to use it or should have introduced them to its use in a rational and then we could say ethical and reasonable way from the point of view of the ethics of legality, are digitally illiterate."

[Interviewee #17] "Well, it can affect I believe fundamentally his freedom to make decisions and his privacy and other fundamental rights that are connected. And in this sense we are linking our decisions and the impact they will have on our society, individuals, collectives, society, democracies, but also the impact this will have on future generations, i.e. we have a responsibility."

[Interviewee #20] "For sure, this is having an impact on how young people relate to each other, like creating friendships, like creating couple or family models. All of this is already being affected by a technology or social networks and that in some way would also use artificial intelligence. So I'm sure it will have an impact on a social organisation."

2.3. Legal domain

2.3.1. The geopolitics of AI

The existing legal frameworks, as well as the new proposals for legislation on AI are configured not only as tools for citizen protection, especially in terms of human rights, but also as instruments of technological competitiveness in a global policy context in which more and more countries are investing heavily in AI. For this reason, the legislative characteristics of a country or region around AI can give rise to numerous opportunities and address challenges, but always from a possible differential fact, which is that not everywhere in the world has (similar) AI legislation. Due to the fact that this may generate a debate around the speeds AI is undertaking in different places, we asked the following question to the interviewees: **in your opinion, what happens when certain types of technological development such as AI are banned or restricted in one country, but not in other countries where AI technological development is a major investment?**

After analysing the responses, we have grouped them into two groups, one with a rather positive or optimistic position and the other rather negative or pessimistic regarding the impact of bans or restrictions on the development of AI and its repercussions at the global level.

POSITIVE POSITION OR OPTIMIST

[Interviewee #1] "One thing is how we generate it and the other is how we receive it. [...] People will end up buying the available technology and if the available technology is China or the United States first, well [...]."

[Interviewee #8] "As a big challenge, a big challenge. In the end it is all part of the work of the European Commission. So I agree with this and I think it is a job that has to be done because it is the one that protects the sea from the possible damage that these misuses can cause. Having said that, of course, when you put this in a global context where there are other actors that double, triple or multiply by several orders of magnitude the investment and they do it from a totally liberal perspective. So here

I do think that we have a very difficult challenge to face because at the moment we are faced with a totally illiterate consumer [...] but I also see a task of how to give the necessary culture to the consumer so that he or she knows how to choose."

[Interviewee #14] "I think we have to do all this in a coordinated way. I mean, I would say we should convince China and the United States to take joint action like we did with the nuclear bombs, so that all of a sudden you don't have autonomous soldiers who have a software bug killing people. Unfortunately I'm sure it already exists."

[Interviewee #15] "We cannot judge the actions of the Chinese from our own perspective. [...] It is reprehensible, but from the Chinese moral, ethical and legal perspective, which is the one in force there, nothing to say."

[Interviewee #17] "If we look at it in terms of economics and competitiveness and to see who wins the race, it is clear that either the fewer restrictions and less control, the more [China] will win, but I do not believe, and I have said it before, that we have to look at ourselves as a mirror in which to reflect ourselves in these dynamics that have nothing to do with our cultural, political and social context. I think that in this sense Europe stands out. It has often been said that it is not a pioneer in innovation, but precisely because we have this basis of respect for values and human rights that should be a priority and a priority."

[Interviewee #22] "Yes... when you compare this we have here the whole artificial intelligence strategy which they say is human-centred. Some people criticise it and say well this may slow down progress, it may slow down development and therefore it may have negative repercussions on the economic development, right? Because China or the United States are, let's say, not worrying about these things and they say they are moving faster. Well, I see also another side of this coin because, let's see, really... There is one thing that is key for me and that is the trust in the artificial intelligence system."

RATHER NEGATIVE POSITION OR PESSIMISTIC

[Interviewee #2] "What worries me is the impact that development in developing countries can have here. In some ways China may have a stronger

entry. But we have to consider that China's authoritarian and bureaucratic political system of values has repercussions according to the development of artificial intelligence itself. So I am not particularly worried about China. I'm more concerned about how artificial intelligence evolves in the most advanced country, which is by far the United States."

[Interviewee #3] "I want Europe to be a regulator of strong ethical principles and I want Europe to have a strong position against artificial surveillance in public spaces and against killer robots. But the problem is that we live in a global world and we don't have walls, we have high walls that can make this space that we want ethical and these principles strong. Because in the end technology interacts with other parts of the world. I think that is a big challenge. I was very struck by a story that came out in the media about the new president Joe Biden, who after two or three weeks in office, one of the first sessions he had was to lobby the artificial intelligence think tank. [...] He asked them to please don't block the development of killer robots. Why? Because that development is happening at high speed in China, it's happening at high speed in Russia, with very, very large budgets, and it puts the United States at a military competitive disadvantage. Well... one thing is what we would like. Another thing is the reality, the geopolitical tensions."

[Interviewee #4] "Well, I think this is going to have a direct impact on the global competitiveness factor. That is to say, I think the fact that China can be a more lax power in terms of restrictions from an ethical or regulatory point of view. [...] It may mean that the advances that take place in the coming years, let's say, will put the other powers such as Europe and the United States at a disadvantage when it comes to competing on a global level."

[Interviewee #5] "Of course, the European view is much more protectionist in terms of social impact and protection, privacy provision and user protection. [...] And that, of course, allows China to go faster technologically."

[Interviewee #7] "Of course, to dominate, to obtain a position of leadership, obviously also implies a greater possibility to lead economically, in short, all the derivatives that this entails, doesn't it? So, of course, this is in contrast to the ethical vision we have in Europe."

[Interviewee #9] "I think Europe is completely alone on the issue of legal protection. The misuses or risks of artificial intelligence are the same as in the protection of data, and at the same time, politically, the two big blocs, if we look at the United States and China, are putting pressure. Neither is strong enough to control from the Chinese side [...] and I see little future in what Europe is trying to do, a very restrictive or very protective law or regulation that is then impossible to comply with."

[Interviewee #11] "Well, as I said, there are two important games here. [Technology moves, is exported and imported. So we have to be very careful in this area, because in the end a country that is less regulated in this area is the one that is the pioneer, because this is also what is happening. In other words, in the end, the places where they have less regulation tend to be more pioneering in this type of activity."

[Interviewee #12] "I was telling you before that technology is neither good nor bad, but it is not neutral. And obviously the fact that it is not neutral also means that it depends on the place, the environment in which you insert it, that is, in societies that are very unequal."

[Interviewee #13] "These decisions are really very high level and have a global impact, because they are determining the reaction in the rest of the global world and macro policy. So, both at the macroeconomic level as well as at the macro political and social level, the impact is very high and at the end of the day the development of artificial intelligence is totally a geopolitical issue."

[Interviewee #19] "Complicated... dangerous..., because of what we said before, we are in a global world. Therefore, if we do not share the global ethic that we mentioned before, if the solutions have to have a local or global perspective, then of course if China uses artificial intelligence to control citizens to a point that is dystopian [...] In China I think it should not be applied, it is dangerous from this point of view, complicated to manage and to limit."

2.3.2. AI governance

Around the world, representatives from industry, governments, academia and civil society are discussing the governance of AI through the development of ethical principles, technical standards and professional codes of conduct to address some of the challenges and opportunities presented by the widespread deployment of AI. While not all of the challenges and opportunities are radically new, many will be due to an unstoppable digital transformation and therefore new approaches and levels of action beyond the conventional ones can be envisaged. Thus, alongside the debate on current voluntary ethical frameworks and technical interpretations, there is a broad debate on where legal and regulatory frameworks for AI are needed. To gather more information on this issue, we asked respondents the following question: **in your opinion, who is or should be responsible for setting and enforcing ethical standards for AI systems?**

After analysing all the responses received, we have made a grouping between people who consider that an eminently administrative governance position is necessary and one in which people are more aligned with a civic or multi-sectoral governance position.

GOVERNANCE POSITION

ADMINISTRATIVE

[Interviewee #3] "I think the digital economy is global. Countries have very little weight in the regulation of technology from a national territory perspective. And here in Europe, well, they have an impact. [...] But I think it has to come out of the United Nations, because the challenge is global and we're talking about well.... 20, 30, 50 years."

[Interviewee #9] "If we look at the European Union alone, there would have to be this regulatory framework that would have to be enforced, as we have seen with the data regulation. [...] If we think of a company or we think of the government or we think of an organisation, yes, there would always have to be a committee. It would have to be independent and it would have to ensure that the ethical values that have been agreed upon by that company or by that government are carried out."

[Interviewee #10] "One thing is to regulate and the other thing is to enforce ethical standards. I think..., but what is clear is that I think it is a question that has to do with public administration. And with the regulatory bodies and therefore this means that whatever regulation it is, it must have

entities that supervise it and then have the capacity to enforce it."

[Interviewee #11] "Well, in the end it has to be of the highest level in order to be a little bit generic. At the European level, it is clear to me that we have it here, we have everything very well, it has worked quite well with the data protection regulation when the European directive came out, so we should follow the same path".

[Interviewee #18] "Well, here I am in favour of regulation at an official, state and supra-state level. So it seems to me that, well, there are certain very tactical issues that could be regulated at the regional or municipal level, for example. But I think that the bigger, more important issues, which affect questions of dignity, questions of justice, questions of access, these questions have to be regulated at the state level."

[Interviewee #19] "If it is a product, it is the company itself. But the one who has to set the limits of the company will be the legal part and there must be a political will. [...] Therefore, I think it is a shared problem and there is no one directly responsible. And what scares me is that there is no agreement to share between the different agents."

GOVERNANCE POSITION

CIVIC

[Interviewee #4] "Let's see, I always say that my approach to the issue is quite open in the sense that I think that different actors have to collaborate here. Of course, the public administration has to establish a basis for what is legal or not legal. But beyond that, I always say that one thing is legal and another thing is ethical, and often something can be legal and not ethical, so I think that private companies should also make an effort to encourage a commitment from top management to develop ethical artificial intelligence solutions. And then I think there also needs to be some pressure."

[Interviewee #6] "How many laws are necessary? I would say it should be the minimum. But this requires a condition of high ethical quality. Let's say that in the most radical and pure anarchist thought it is clear that if everyone were ethically very good, there would be no need for legislation or police. [...] We will have to ask the public power to step in to guarantee the minimum legal path. [...] Therefore, maximum ethical quality minimum

legislation, minimum ethical quality maximum legislation."

[Interviewee #12] "I think that society should take responsibility for that. [...] Citizens and users in general. Why? Because I wouldn't expect anything from anyone. I think it is good that we are belligerent, that we are committed citizens and that this is our responsibility. That obviously through our mobilisation we have to put pressure on governments. Because they are the ones who design public policies. Why? So that these public policies are consistent with the protection of people's fundamental rights. But I think it's our responsibility as citizens."

2.3.3. The regulation of AI

Alongside AI governance, there is a growing focus on AI regulation to set the boundaries for action to ensure the safety and performance of commercialised AI solutions. Regardless of the contextual differences and potential loopholes posed by some AI solutions, we know that a large part of standards are led by the private sector, not only around ethics and standars, also with regard to AI regulation itself. However, it can be said that AI regulation is still in its infancy as shown by the first regulatory actions or proposals and statements from governments and AI agencies around the world. While some legal issues such as data protection and privacy enjoy a more advanced trajectory, other issues such as transparency, surveillance and accountability or human oversight are at a less advanced stage. Recognising the importance of how different regulatory initiatives may shape up around the world and that different approaches will be needed to ensure the safety and performance of commercialised AI solutions, we asked respondents the following question: **which type of AI regulation do you consider more appropriate today - restrictive (the regulate and forget type) or adaptive (the iterative with technological change type)?**

Taking into account the responses obtained, we have grouped them into three groups. Those that reflect an adaptive position, those that denote an adaptive but at the same time restrictive position, and those that could be described as proactive.

ADAPTIVE POSITION

[Interviewee #1]

"I think it has to be adaptive necessarily because technology works so fast that standards become obsolete very quickly."

[Interviewee #4] "I defend the adaptive approach, more than anything else because I find it very difficult to think that with what it costs to launch a regulation in an environment such as the European one, we are going to be able to cover, well, the advance of technology. [...] I think that here we have, as I said, to combine this capacity to innovate with adapting ourselves from the point of view of regulation and social awareness so that later, as I said, it is the citizens themselves who are very demanding with the companies. In the end, rather than prohibiting, I think it has to be a matter of ensuring that certain products and services that stand out as unethical do not have consumers and, therefore, do not have any kind of financial viability."

[Interviewee #9] "The adaptive part I do like more, much more because the technological changes are very fast and it should be like that. But I'm afraid the analysis is not done how it should be dealth with and how it should be nuanced so everything is very broad, which means that margins are not set [...]. It is increasingly clear to me that the limits of how far you can go are important, extremely important."

[Interviewee #19] "The adaptive and operational always because the world evolves."

[Interviewee #22] "Well, my common sense, my intuition, tells me that it has to be more adaptive, but what is clear is that it is changing almost on a daily basis. Well, I don't think it would make much sense to make something... what do you call it... carved in stone that can't be changed. [...] So clearly it has to be done in a much more intelligent and adaptive way. This is crystal clear."

ADAPTIVE AND RESTRICTIVE POSITION

[Interviewee #3] "I would say it has to be a mixture of the two. There are a number of red lines that perhaps have to be restrictive, but always from a perspective that in the digital economy everything that touches technology, there is no possibility of legislating ...only in new areas, you know you have to touch again. [...] The problem sometimes are the times of consensus to regulate that sometimes are not the times of technological development [...] I think that from that perspective it is quite positive that at least Europe is putting in place a piece of regulatory framework that has never been regulated."

[Interviewee #5] "It's an excellent question and I don't think I have a closed answer here first, OK? I have thought a lot about this together with other colleagues and I myself do not have a closed answer, but it is very clear to me. Only a punitive vision will not work. It won't work because this punitive view assumes that the responsibility lies only with the producers and I don't agree where everyone's roles are distributed. Yes. Why? Because there is an element of distributed decision-making."

[Interviewee #6] "As long as adaptive does not mean submission. Adaptation is one thing and submission is another. Therefore, let's say agreed adaptation, let's say that we have to know how to create the conditions of a certain social contract and in the spirit of the contract there is a degree of mutual adaptation, i.e. of the different contracted agents".

[Interviewee #15] "No, not a chance. We have to change the economic model. We have to do a reset. Oh, and we have to give people their privacy back and start from scratch. Adapt."

PROACTIVE POSITION

[Interviewee #2] "I believe that neither of the two, it has to be proactive, that is to say reactive without epic possession of politicians in the face of possible dangers. [...] I believe that we are now fundamentally entering a society of responsibility and [...] as that philosopher said that it is ethics and responsibility that must make people aware that they have to be responsible, and it is more of an educational process than a legislative one."

[Interviewee #18] "For certain technologies that are absolutely immature, such as facial recognition, posture analysis, gait analysis, all those kinds of things, [...] I am in favour of a moratorium. Now, a moratorium implies that it is not a ban. [...] It can be taken with a de facto moratorium. They are not authorised."

2.3.4. The social justice of AI

We know that the use of automated decision-making processes involving AI systems can affect individual people in a differentiated way, and also society at large from a social justice perspective. There are more than a few cases where the use of AI has had a social impact in terms of inequality, social ranking and social division. While AI is receiving unprecedented attention due to its presence in multiple spheres of our daily lives, the implications it may have for social justice and human rights are still understudied, even knowing that algorithmic decision-making, especially for marginalised and poor people, can undermine social cohesion and social justice. Thus, taking into account a vulnerability perspective we can focus on how the use of AI systems also builds relationships between individuals and institutions, and such relationships give the possibility for these systems to affect people positively or negatively. In that sense only the design of AI decision-making processes will largely determine whether their use results in greater cohesion or segregation between individuals. To gather specific information on this issue we asked respondents the following question**: in your opinion, how can we ensure that the algorithms used in AI systems are fair, especially when they are privately owned by corporations and not accessible for public control?**

Following the collection of information we have grouped the responses into two categories, those that reflect a rather optimistic positioning (e.g. through certification) and those that are rather pessimistic, as they consider that we cannot make the algorithms used fair.

OPTIMISTIC POSITION

[Interviewee #1] "What I think is what I was saying about the quality seal or the ethical seal that some and part of the administration have undertaken to certify that this product or this company works with appropriate ethical codes. The consumer already has a guarantee in this and thus offer what we were saying before."

[Interviewee #3] "Perhaps what we should be asking ourselves is what mechanisms we can have so that when necessary the public or an independent authority can audit systems. [...] We audit all sorts of things that companies do, don't we?."

[Interviewee #6] "As long as there was sufficient awareness and ethical quality in the processes of creating artificial intelligence we would not have to

make legislation to control this justice, but as long as we do not have intentions of good ethical quality of the whole artificial intelligence production apparatus it is urgent to legislate, i.e. it has to be done."

[Interviewee #7] "I don't know any way, but I think that through this system of public regulation and therefore in what has to be enforceable, we should try to find a way to make this company accountable for adjusting its algorithms."

[Interviewee #8] "With the certifications, the algorithm has to certify that the algorithm is fair. So the algorithm does not have a biased design and has to certify that it is used with certified training data."

[Interviewee #14] "Actually there are two ways depending on whether you are in an Anglo-Saxon country or not. In an Anglo-Saxon country you would say, well I trust you and then ask for accountability. In the rest of the world, in the rest of Europe? [...] We are going to have to ask for transparency and a final audit. But if you look at it, they are completely the opposite. [...] I know I have to ask for transparency and if you don't give me enough transparency, we're going to audit you because I don't trust you."

[Interviewee #15] "You can't. What you have to make sure is that all the ones that the government uses, that the government licenses are. Otherwise I would tell you that if I had the power to legislate I would ask that before they could be launched on the market, any product, I don't need to see it, but there might be something that would certify them and like the clothes we all wear, they would have one of these labels that says washable and there would be a traffic light, 4 or 5 icons: green, red, green, green, yellow, red. That would be enough."

[Interviewee #17] "This forces us to move towards new models of governance or public and private initiative. [...] Of course this is a paradigm shift. But I am not in favour of the fact that there is very cutting-edge private initiative or that states have a certain level of control and that this opacity or non-transparency is the general dynamic. I am totally against."

PESSIMISTIC POSITION

[Interviewee #2] "I think this is impossible. It's impossible from the external point of view of the company itself."

[Interviewee #9] "I think that if they are privately owned by corporations and are not accessible to public control it is very difficult to ensure that these algorithms are fair. [There is no transparency here, there is no accountability. There is no feedback to know what is going on. So I think there is no way to ensure that they are fair."

[Interviewee #12] "We can't. We can't. In the past, in a capitalist system where things are patented and where there is private ownership of what is patented, you cannot guarantee that."

2.3.5. Transparency in AI

There is some consensus that in order to gain trustworthiness in technology in general and AI in particular, transparency must be improved, especially when the use of algorithms can have significant effects when it comes to automated and important decisions about individuals. In Europe, this debate is largely focused on so-called algorithmic transparency and accountability, and is often related to compliance with the so-called right to explanation, enshrined in the EU's General Data Protection Regulation (GDPR). However, transparency is not only about explainability but also about interpretability and trust, i.e. how ordinary people understand explanations and how they evaluate them in light of the fact that there is an AI system providing or facilitating a service or product. Of particular interest here is the relationship between transparency and trust, which are key objectives for the European AI strategy since the publication of the European Commission's White Paper, where the importance of the "ecosystem of trust" is clearly underlined (2020: 9). However, it should be stressed that this is more a goal than a generalised reality as in the vast majority of AI models we derive the functionality of the model from some data and through the use of algorithms that attempt to build the most accurate model, but not necessarily through the most transparent model. Hence the terminology of "black box". It is in this sense that we posed the following question to the interviewees: **in your opinion, how do we balance the need for more accurate algorithms in AI systems with the need for transparency towards the people affected by these algorithms?**

The responses obtained reflect two groups. First, people who favour accuracy over transparency, and then people who favour both accuracy and transparency (although the order can be reversed).

POSITION THAT FAVOURS ACCURACY

[Interviewee #7] "I think it depends a little bit on the application. I don't know if I'm talking about applications, for example, in the field of clinical diagnostics. Then I would probably be interested in algorithms that are very efficient, right? [...] Accuracy is more important than transparency in the algorithm. In applications where there are, let's say, there may be implications in terms of decision-making, which may lead to potential discrimination or this kind of thing, then obviously the transparency of the results has to be more important to me."

[Interviewee #8] "Between precise and safe or transparent, precise and safe. Of course. [...] I think there is also a debate about transparency. I think there is a point of confusion. Everybody is assuming that being transparent means revealing the algorithm. And I think that this is not the case. In other words, in order to be transparent you have to be able to explain the criteria according to which the algorithm decides, or you have to be able to explain to a given input with which criteria the solution is elaborated or according to which principles the solution is elaborated."

[Interviewee #11] "Let's see, I would stay with precision and security."

[Interviewee #18] "If it's just on a conceptual level, I'd go with the first one. The thing is that the security thing.... It's very difficult, assuming that this security is real and for everyone."

POSITION THAT FAVOURS ACCURACY AND TRANSPARENCY

[Interviewee #3] "Well, I think that accurate and secure must always be the first point of departure. I mean, we cannot afford to have algorithms that are nothing else. For me, they are like primary elements and transparency has to be proportional. I mean, maybe as citizens it's not that I need all algorithms to be transparent, I need tools [...] that don't come to me and say that you can't access or audit an algorithm because it is protected by intellectual property. When we know that this algorithm has a social impact, it has an impact on society and the common good for the welfare of society. So transparency is important."

[Interviewee #4] "I start from the premise that there are algorithms that have a very high level of explainability, for example, a regression, because it has a perfect explainability, but that does not imply that it has a very high level of precision, which means. [That is to say, if with a regression I am able to give a result with a very high level of precision, then obviously if I am going to opt for this, I don't know if I want to use a neural network, which in the end is much more of a black box."

[Interviewee #6] "Transparency is not incompatible with precision and security. [...] We will find the maximum of precision and security, but always transparent."

[Interviewee #12] "Well, I think there can be no discussion. Algorithms have to be transparent. Full stop. In other words, an algorithm that is not transparent is not even useful because we know that there are algorithms that give us results that we don't know why they have given us them. So, if we are not able to understand that. And the lack of transparency obviously contributes to this, this algorithm is not useful. In other words, one has to be understandable, one has to be transparent, and anything less than that doesn't interest me."

[Interviewee #13] "Of course, I have it easy here because I advocate for a more supervised decentralised technology and that would be everything that is the development of blockchain, which also guarantees transparency because it provides a higher layer of security. [...] But the solutions, those solutions, one is secure, accurate and transparent technology. They cannot go separate ways. They have to be one solution."

[Interviewee #15] "Algorithms, by definition, have to be secure and transparent. I mean, surely if you ask on the street, another answer would be permissible, but I am a professor at the Faculty of Computer Science and we are interested in the accuracy of the algorithm and because we are discussing ethical issues, we are interested in them being transparent, there is no doubt about it."

[Interviewee #17] "Not having access to both I had never imagined."

[Interviewee #19] "It should not be incompatible. This is good to start with. [...] There are some systems that we are able to explain in the background what

system of equations is behind that models this learning. Now, if it is an aid system I would like it to be transparent."

[Interviewee #22]

"I don't see a dichotomy here. Why does it have to be one thing.... What we have to do is make very precise and transparent algorithms. Of course, in today's Deep Learning techniques, that's fine. [But we don't have to go this way. We have to develop new artificial intelligence systems, new techniques that allow us to have precision, quality and transparency at the same time. I see it as perfectly feasible".

2.4. The future outlook

2.4.1. The main ethical and social challenges in the long term

Although the future remains uncertain regarding the development of AI, there is a consensus that the widespread implementation of various Machine Learning techniques has been a giant leap in the more recent course of AI, and especially in reference to its development in the area of perception resolution. In this sense, some current AI systems take information from visual, auditory or speech input directly and, in some cases, no longer require human supervision. In the future, such developments are likely to become increasingly pervasive and will raise ethical questions related to the possible existence of systems or machines capable of performing tasks autonomously and superior to those of humans. While people have worked together with machines, using them to make us more productive and efficient, massive data and more powerful algorithms are beginning to change the landscape and the relationship between people and machines, to the point where machines can learn and improve autonomously.

It is for this reason that ethical issues related to a possible general AI raise different ethical questions than those arising from an actual use of information-based AI and process automation. Thus, we can say that ethical issues could become more frequent as we give more power to a machine or robot. Thus, an ethical problem might be, for example, to support or oppose the development of lethal autonomous weapons systems on the basis of whether they increase efficiency and minimise civilian casualties or, on the contrary, allow terrorist groups to take control of certain military conflicts and violate

fundamental principles of human dignity. Another frequently discussed problem is the potential for AI to spread fake news, psychologically manipulate people through targeted emotional appeals, and even hide dissent through armies of bots. While we have always been vulnerable to being duped, provoked or manipulated, the use of AI facilitates unprecedented scale change and immediacy through algorithms that check content over and over again across millions of people and at great speed. With this in mind, we posed the following question to respondents: **looking at a long-term (25-year) future trajectory, perhaps in a general AI context, what do you think will be the major ethical and social challenges that will cause governments to oversee, shut down and/or nationalise AI systems?**

Taking into account the opinions received, we have made a double grouping that highlights, on the one hand, the position on geopolitics and the development of AI, and on the other hand, the position on the social impact of AI. Many of the responses also reflected uncertainty about the long-term future of AI.

POSITION ON GEOPOLITICS AND AI DEVELOPMENT

[Interviewee #4] "I always say that I avoid reproducing Black Mirror-type schemes. That is to say, I am not one of those who think that the future in 25 years' time is going to be a dystopian future, where there are no governments and companies and large corporations have taken control. [...] Having said that, I would point out that Western societies fortunately do not tend towards this model of nationalisation of private resources. But going back to what we were saying about China and the great powers, I think that what is going to change here, or should change, is public-private collaboration."

[Interviewee #9] "The first challenge for the next few years is to get us to agree. [...] We are already doing it in Europe, but we do not agree on how to apply these ethical principles. [...] If we agree on how to apply these ethical principles, whether by regulation, by law or by ethical guidelines, and we start to apply them..., the other challenge is social so that people, society in general, without knowledge and without the obligation to know the technology, can understand the impact that the current technology, in this case artificial intelligence, has on their lives, and have the capacity to decide where they set the limits in their lives."

[Interviewee #14] "Well, I think China already does it to some extent, so I know it's not future, it's present and who knows what North Korea is like, for example, we don't know, but it could be worse. I think it's a very difficult question. I liked the 25 years. It's a good number. [...] I think there is only one challenge that has to do with everything we have talked about, and the challenge is what is the balance between regulation and freedom, let's say, of technological development."

[Interviewee #15] "Let's see, I don't think it will happen in the United States .... I don't think it's possible in neoliberal countries, and in capitalist countries, where research and digital industry go hand in hand. There is no need to nationalise anything. And besides, there is no money to nationalise them. So there are other countries that have chosen a different path. This is all national, party-based. China, India, Singapore. [...] The biggest challenge is to re-educate people in the sense that they know the potential we have and the challenges we face."

POSITION ON THE SOCIAL IMPACT OF AI

[Interviewee #3] "Well, I'd like to think that we have more than 25 years to go before we get to that general artificial intelligence, because I think we have a long way to go as a society to be ready to deal with that reality. We have a hard time. [...] My first question is, where is this technology going to develop?."

[Interviewee #5] "Look, I think the first thing is an impact that we should already be working on, okay? First of all, let's start at the grassroots. There are different open source and Creative Commons initiatives and so on that are trying to open up what would be the equivalent of Wikipedia or Linux or free software to artificial intelligence frameworks. [...] This is fundamental. It is fundamental, above all in order to be able to develop an artificial intelligence commons. The commons is the reference point. And I don't think we are devoting enough effort to this commons. Why not? Because you have said it very well, because when the market moves forward, it moves forward."

[Interviewee #6] "They have already been anticipated. I would say that the main challenges will be, because we are talking about the main ones, the cure and the radical respect for privacy and freedom. [...] As we understand

it in our personal autonomy, which is partly made up of privacy and intimacy and which has a maximum expression of freedom, the loss of autonomy, which was the great conquest of the Enlightenment, will be marginalised."

[Interviewee #8] "I think I don't have so much perspective and I see so many challenges now that I don't know if I am able to imagine. I mean, I would like to imagine a future in which we have been able to articulate a future in which we have been able to articulate all this ethical use of artificial intelligences where these artificial intelligences are occupying a social space which is to carry out the most unpleasant tasks for humans or to increase the capacity to solve the most complex problems in shorter times and with better quality and to provide accompaniment. Of course, there are many risks."

[Interviewee #10] "The main challenges are the same now 250 years ago and 25 years from now, with the difference being the scale, the magnitude and the capacity that these artificial intelligence mechanisms can bring. [...] We are kind of confident that when we choose the variables we take into account to make certain decisions we are not creating disadvantages between one social group and another."

[Interviewee #16] "I think we will be faced with functional challenges. They were things that machines will be able to do but that people will want to continue to do. Everything will be much more hybrid than it seems. [But I think we have a bias that we talk a lot about the progress of artificial intelligence and very little about the progress of human intelligence. So I'm not one of those who thinks that this is an impossible thing to manage."

[Interviewee #21] "Two issues in the final part of the question, which is whether to close or nationalise. The truth is that it can be regulated, but I don't think you can put gates on the field. I mean, I find nationalisation difficult. I can't imagine a 1984. I think that the development of these tools will hardly be monopolised by the General Staff. Now, with regard to what you said in the first part of the question. I think what is going to happen is the same thing that has happened with regulations in other fields such as those I mentioned. Social outrage is the main driver. I mean, if we don't move, this will be left to private enterprise and we will suffer the consequences. The general population."

[Interviewee #22] "In fact, this look to the future is not the one that worries us most because we want to focus on the present and looking to the future means that we leave the present. But we have to worry about something if we don't. [...] For me the problem is the degree of autonomy you give to a system. [...] Maybe we should bear in mind and see what can be done to avoid possible very negative repercussions of decisions taken by very autonomous algorithms that could affect us."

2.4.2. The balance of opportunities and risks of AI in the future

AI is often presented as a catalyst to accelerate technological progress while providing mechanisms to overcome traditional information analysis and management obstacles for multiple sectors of the economy in a cross-cutting manner. But beyond the technical challenges, the implementation of AI on a widespread basis means dealing with many challenges that originate from existing social conditions. In that sense, although it is becoming increasingly common for computers to perform some tasks better than the best humans and often provide valuable information for dynamic risk assessment, not all tasks are amenable to automation, nor does such automation guarantee the best or most sustainable solution. We are aware of the problematic developments regarding the combination of complicated social concepts with simple statistics in the field of AI, and given its broad impact on sectors such as finance, education, criminal justice or social welfare, it is to be expected that many complex and pressing issues can only be successfully addressed from a multidisciplinary perspective. With this in mind, we asked respondents the following question: **in your field, do you think that the advantages or opportunities for AI development outweigh the major drawbacks or risks in ethical and social terms?**

Based on the opinions received, we have grouped the responses into two positions, a positive position which would reflect more advantages than disadvantages and a negative position which would denote the opposite.

POSITIVE POSITION

[Interviewee #1] "I am an advocate of technology and therefore I believe that the advantages have to outweigh the risks by far. It's not that the risks don't have to be prevented and have to be mitigated and I hope. Well this is the attempt. [...] And it is clear that they have obvious biases that

we are often not aware of. It's not that we want to do it this way, but the circumstances make it develop this way. [...] It's very important that women get involved in the development of these technologies."

[Interviewee #2] "I think so. At the moment, everything must be in a state of paralysis.[...] Technology is a manifestation of human freedom, a cognitive system that is all the time projecting a future of new things forward. We humans are the parents of technology. We are the children of nature, but we are the parents of technology. [...] So far it has worked well, we have developed formats and we have created the ecosystems that somehow artificial systems work. Now we are at a very critical moment. Effectively this civilisation is reaching a moment of crisis and rethinking. And this is where we have to somehow situate these opportunities or doubts that artificial intelligence can bring to the table."

[Interviewee #3] "First, it will be to demystify artificial intelligence so that we as a society can really get down and make the Black Box disappear, this whole black box that intimidates us so that we can talk about it. Because without the automation of the economy ... [...] But we have to talk about what happens to jobs, what model is given. It will be a bit like nuclear energy, won't it? What do we want from this technical potential? What do we want to do as a common, as a society?"

[Interviewee #4] "Look, I would say yes. I mean, I think that right now the opportunities that we have with AI in the area that I work in, which is helping our clients to improve their products and services, to be more efficient. There is still a long way to go before we reach a level of, let's say, high sophistication that makes the social issue, the ethical issue very relevant."

[Interviewee #8] "I think it is so important that we do this great work to raise awareness of the desire for ethical development ... [...] So I think that as a community, I think we also have enough intelligence not to collapse our own survival as a collective. [...] But I do think that it depends a lot on the two legs of wanting to develop the idea of this beneficial objective and wanting to use the line only for this beneficial objective."

[Interviewee #10] "I think it necessarily has to be. It has to be that way because, if we don't have problems accepting the social ability that this, ultimately,

it ends up affecting trust. [...] I do think that the trust of citizens, I was going to say users, but more and more we are talking about users in the end if we want to stand up for rights we have to think in terms of citizenship. I think that trust is a fundamental value."

[Interviewee #13] "The answer is yes, isn't it? I dedicate myself to what I do, precisely because I am convinced and I believe that this is what information technologies, those of us who are and come from the world of engineering, and who also have management skills, are for. What we must do is provide everyone with this opportunity. In other words, power. It really should be to have much more democratic principles at the level of accessibility."

[Interviewee #15] "Yes, yes, I think so. I think it should be ... Uncontrolled and unregulated development has made us very quickly aware of the risks. It has already systematically obscured the advantages. And now we are in a period of self-flagellation as a community and we are explaining to everyone what has gone wrong and what is risky. Because surely those of us who have reached a certain age and a certain position in the community realise that those who come after us are not realising what is going on. And that may be the last thing we do. But you think about the average level of people in the ethics forums, and with some exceptions, very young people."

[Interviewee #16] "Radically yes, I think it will have an extraordinary impact. Artificial intelligence plays with all the data technologies I think will have an extraordinary impact and I think it will give us a lot of opportunities that we have. And I think our challenge is to make the sum of intelligences ... [...] I don't think it's that easy to replace people. I really like a question at the beginning that defined well those things that people knew how to do."

[Interviewee #17] "If we achieve a balance, but clearly it is up to us and we have this opportunity, which is a magnificent opportunity to do things right and that is why we have to respect that this ethic that we have built, that we have agreed on or that we are starting to build, and take this perspective has to respect human rights."

[Interviewee #21] "I think so. This doesn't mean that we don't have to be very attentive

to the biases that are produced. [...] What is good is to see what we do, how we take on this problem, which is serious. I mean, besides, ethics in itself is also a somewhat changing thing."

[Interviewee #22] "Well I am of an optimistic nature here, and I would say that if there is no bad judge I think there is. I think the potential for positive and socially responsible applications is very high and it's very clear to me. It's clear now that it will go where it will go."

NEGATIVE OR SCEPTICAL POSITION

[Interviewee #6] "I had already predefined myself as an optimist before and I tend to believe that I am. However, if someone asks me that question, a percentage. [...] As a Catalan philosopher of the 19th century said that he was moderately sceptical within pessimism, I am moderately sceptical within optimism."

[Interviewee #9] "I would like it to be like that but I can't say because there are many variables, there are many factors ... [...] But allow me not to be so utopian and to question it a little bit."

[Interviewee #12] "I think it's an open question.... That question is open and as I said before I think it is a society that unfortunately we are evolving towards the worst as a society. So, given that as a society we are evolving towards the worst, my fear is that artificial intelligence will also evolve towards the worst. But it is true that I am still optimistic, I think. [I think we have the capacity to fight for a better world."

[Interviewee #14] "I'll give you a historical parallel. I think this is going to be as difficult or more difficult than nuclear energy, so it's not going to be easy. I mean, think that in the nuclear energy issue, thousands of people had to die before someone realised the ethics of using it as a weapon of war. It took years for that to happen. [...] So the same thing can happen here without end. If we don't learn from history, as we never learn from history, unfortunately if we don't learn from that it's going to be difficult."

2.5. By way of conclusion to the second part

In this second part we have collected different reflections and positions of the 23 people interviewed in relation to three areas of interest on the development of AI: (1) the ethical and social domain, (2) the legal domain, and (3) looking into the future. There is no doubt that this qualitative data collection exercise has been very useful to gain a better understanding of how different morally challenging circumstances proliferate and are interpreted in contemporary AI-related societies. Adopting a "practical knowledge" approach addressed through Aristotle's notion of phronesis and the heuristic paradigm, where the notion of interdiscursivity plays a central role, we have analysed the contextuality of the information from the interviews with 23 people from different professional and academic backgrounds, capturing a good part of the (very real) difficulties and opportunities for AI progress.

In the first part, within the ethical and social domain, we have collected positions of the people interviewed that have to do with both the design of AI and its social impact. In the first section, we have seen how there are different perspectives on whether ethical considerations in AI should be a restriction, a sub-objective or the main objective, mainly in terms of AI design. In this sense, the reflections are primarily placed on giving greater importance and significance to ethical considerations in AI as evidenced by the fact that the positions oscillate between being a main objective and/or a configurative part of AI. However, we also find positions that are rather neutral, and partly expectant, willing to let the development of AI itself set its limits and ethical considerations. As to whether artificial intelligence could be a factor of human weakening or strengthening, most of the discourses tend to point to a vision that could be considered optimistic but with reservations. Thus, on several occasions it is envisaged that AI may provide an increase in our capabilities, even autonomy for many people in an ageing future, but there is also the fear that we may lose certain physical and cognitive abilities in this process of change, and that AI may dehumanise us.

Regarding the importance of the context of the ethical considerations of AI, a large part of the positions are strongly in favour of taking actions that would allow a better understanding of the context by including diverse opinions, situations and geographical spaces. Along these lines, the idea is put forward that it is not possible to innovate without carrying out a critical reflection on the AI innovation process itself, which means

taking into consideration not only the people involved or affected in order to reason the best possible solution in each case of AI application, but also broadening, reaching a consensus and universalising criteria and solutions so that they have a greater impact. Finally, this first part also asked whether the younger generations could be particularly affected by the widespread use of AI systems. On this, several positions see the future of the younger generations with regard to AI in a negative way, especially with regard to their freedom to make decisions and also their lack of privacy or intimacy. They underline that there is a lot of work to be done to move towards greater digital literacy, including awareness and skills of what algorithms represent and the consequences of their application. On the other hand, the need for younger generations to think critically about AI is highlighted so that they can elucidate both the benefits and the risks. In addition, several respondents raised the need for regulation to allow these younger generations to use and develop AI technology in a protective framework.

In order to obtain complementary information to the open-ended interview questions, in this first part we have also addressed questions on different ethical and social considerations through closed questions (using a Likert scale of 1-5), generally introduced as stimulating or challenging. In the first question we have addressed a common theme within the more philosophical study of technology, which is to consider whether people are not computer processes or programmes, but unique with empathy, self-determination, unpredictability, intuition and creativity, and therefore have a higher status than machines. It is worth noting that most of the people interviewed agreed or strongly agreed with this statement, and only a few took an equidistant position (neither disagreeing nor agreeing). We also asked whether a wide range of actions, resources and opportunities are being considered to increase the potential benefits and minimise the risks for the younger generation in the face of widespread implementation of AI systems. The responses indicate a tendency for respondents to disagree with this statement, although there is also a not insignificant group of respondents who did not answer specifically (due to lack of knowledge or simply not wanting to answer).

In the second part, within the legal domain, we have dealt with several issues interrelated around the geopolitics of AI, its governance, regulation and social justice. On the first aspect, the positions are rather negative and denote a certain pessimism about Europe's current position in relation to the global commercial and military AI race, with the United States of America and China far ahead of the rest. While there is majority support for Europe to be a leader in responsible AI and a regulator of ethical principles in the AI industry, there is concern that it stands alone in these terms and

is unable to lead competitively and/or strategically at a global level, particularly given China's rapid progress. In this section we will also ask about AI governance and, more specifically, who should be responsible for setting and enforcing ethical standards for AI systems globally. There are two positions here; one that would denote the need for administrative governance at different levels (e.g. from the United Nations to local councils), and another that would indicate greater support for civic governance, in which it is society itself that becomes increasingly co-responsible for AI developments. It should be stressed that the latter is proposed as complementary to the former and, as some reflections point out, requires a high ethical quality from each and every one of the individuals who make up society.

As for the type of AI regulation, there is some consensus that it should be adaptive rather than restrictive, although there is also an intermediate position that it should be a mixture of the two or even proactive. Several interviewees consider that adaptive regulation is inevitable in the face of rapid AI developments, although they do not dismiss that it should be effectively regulated. At the same time, it is stressed that in the face of the development of AI systems that may be considered immature, inappropriate or even discriminatory (e.g. facial recognition, sentiment analysis, posture and gait analysis), there may be a moratorium or prohibition depending on their use. To determine respondents' positions on social justice, we asked how we can ensure that the algorithms used in AI systems are fair, especially when they are privately owned by corporations and not accessible to public scrutiny. On this issue, there is some pessimism that (mainly) the private sector or companies agree to be fully transparent in the development and use of AI systems. However, most are convinced that transparency is a basic principle and the way forward for better accountability of AI technology. In this line, they underline the importance of a governance model in which internal and external control mechanisms are in place to ensure that AI is not only beneficial on a personal level, but also that an assessment is made of the impact on people with less agency, marginalised and poor to maintain social cohesion.

On the issue of whether it is possible for algorithms used in AI systems to be fair, especially when they are privately owned by corporations and not accessible for public scrutiny, the positions are mostly optimistic and denote the importance of developing AI systems in which a principle of social justice can be integrated. To this end, it is suggested that it is most feasible to use certification that the data and algorithm design is unbiased and that there is the possibility of accountability or auditing of the AI system to ensure this. Although few people are pessimistic on this issue, some positions highlight relevant aspects, such as the difficulty of auditing proprietary

algorithms when there is private ownership. Although such a situation would make it difficult to evaluate, it is to be expected that the impact of the algorithm, not only its design, can always be audited. In the legal section we also asked how we can balance the need for more accurate algorithms in AI systems with the need for transparency towards the people affected by these algorithms. The majority of respondents favoured both accuracy and transparency, with only a few favouring accuracy over transparency. However, it is worth noting that the latter group stresses that, in the face of certain uses and potential discrimination, transparency is essential.

Again, in order to obtain complementary information to the open-ended interview questions, in this second part we have also addressed questions on different aspects related to the legal field through closed questions. In the first question of this type, we asked whether the responsibility for an IA decision, action and process should always be assumed by a natural or legal person. Not only did the vast majority of respondents strongly agree with this statement, but none disagreed. In the same vein, we obtained a similar response to the question of whether sustainable processing of personal data has to ensure accountability in the short, medium and long term, with a majority of respondents indicating that they strongly agree. In this section we also asked whether the use of biometric surveillance technologies (e.g. remote facial recognition) used indiscriminately or arbitrarily in publicly accessible spaces represents a violation of the fundamental rights and freedoms of individuals. In this sense, most of the people interviewed agreed or fully agreed with this statement. Finally, we asked about the position regarding AI predictive models on crime that over-represent poor, working class, racialised and migrant communities, and whether this represents a violation of people's fundamental rights and freedoms. In general, the responses obtained indicate that respondents agree or strongly agree with the problem raised, and in no case was a contrary response obtained.

In the third and final part, we focused on the future perspective by asking respondents two related questions about the main ethical and social challenges of AI in the long term, and also about their position in relation to a balance of opportunities and risks of AI in the future. On the main challenges, positions that could be inserted in the field of geopolitics and AI development were highlighted on the one hand, and on the other hand, positions more related to the social impact of AI at a general level. In the first case, such positions are not surprising, as the question asked whether major ethical and social challenges could lead governments to oversee, shut down and/or nationalise AI systems. In this sense, it is worth noting the position that neither the capacity nor the intentionality of governments to take such actions is thought possible (especially the second and third), although the appropriateness of agreement between different

countries to address challenges and possible impacts is mentioned (which would include the first oversight action). In the second case, the positions are more related to the need to develop an AI based on social commons and the importance of functional challenges and our autonomy. In this section, it is explicitly and implicitly underlined that AI should not be used as a substitution tool, but should be implemented to increase human capacity to solve complex problems in shorter times and with better quality.

As to whether the advantages or opportunities of AI development are believed to outweigh the major drawbacks or risks in ethical and social terms, the more positive positions that would reflect more advantages than disadvantages are generally in the majority, although there are also rather negative positions that would denote the opposite. In some cases, even among the most optimistic people, the notion that the unethical development of AI technology and the lack of regulation has made it clear that there are potential risks of uncontrolled AI in the future. Even among the more optimistic positions, it is pointed out that there is still a long way to go before a level of AI deployment is reached where ethical and social considerations are much more relevant than they are today. However, it is underlined that this development is fundamental and, at the same time, it is emphasised that this perspective has to respect human rights. The interviewees that were more optimists also take the view that as we move towards the implementation of AI, we will have to think much more deeply about the relationship between people and machines. On this issue, it is pointed out that it will not be so easy to replace people and that, as a community, we should be intelligent enough not to collapse our own survival as a collective. Among the more negative or sceptical positions is the reflection that it should not be a question of AI but of how society evolves. It is also considered that if society evolves for the worse, it is unlikely that AI will evolve in a positive way. In any case, it is assumed that there are many variables or factors that can influence a rather positive development of AI. In this sense, the historical parallel is made that it could be as difficult or more difficult than it was with nuclear power.

Bibliography

Ahrweiler, P., Gilbert, N., Schrempf, B., Grimpe, B., & Jirotka, M. (2019). The role of civil society organisations in European responsible research and innovation, Journal of Responsible Innovation, 6(1): 25-49.

AI FORA (2021). Artificial Intelligence for Assessment, https://www.ai-fora.de/ (accessed 27 August 2021).

AIEI Group (2020). AI Ethics Impact Group: From Principles to Practice – An interdisciplinary framework to operationalise AI ethics, https://www.ai-ethics-impact.org/en (accessed 27 August 2021).

AI4EU (2021). Ethics: Promoting European ethical, legal, cultural and socio-economic values for AI, https://www.ai4europe.eu/ethics (accessed 27 August 2021).

AlgorithmWatch (2020). AI Ethics Guidelines Global Inventory, https://inventory.algorithmwatch.org/ (accessed 27 August 2021).

Amoore, L. (2013). The Politics of Possibility: Risk and Security Beyond Probability. New York: Duke University Press.

Amoore, L. (2020). Cloud Ethics: Algorithms and the Attributes of Ourselves and Others. New York: Duke University Press.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 27 August 2021).

Annoni, A., Benczur, P., Bertoldi, P., et al. (2018). Artificial Intelligence: A European Perspective. Luxembourg: Office of the European Union, https://publications.jrc.ec.europa.eu/repository/handle/JRC113826 (accessed 27 August 2021).

Appen (2020). How to Reduce Bias in AI with a Focus on Training Data, https://appen.com/blog/how-to-reduce-bias-in-ai/ (accessed 27 August 2021).

Asimov, I. (1942). Runaround: A Short Story. New York: Street and Smith.

Asimov, I. (1986). Robots and Empire: The Classic Robot Novel. New York: HarperCollins.

Autoritat Catalana de Protecció de Dades (2020). Intel·ligència Artificial – Decisions Automatitzades a Catalunya. Barcelona: Autoritat Catalana de Protecció de Dades, https://apdcat.gencat.cat/ca/documentacio/intelligencia_artificial/ (accessed 27 August 2021).

Baert, P., & Morgan, M. (2018). A performative framework for the study of intellectuals. European Journal of Social Theory, 21(3): 322-339.

Baeza-Yates, R. (2018). Bias on the Web, Communications of the ACM, 61(6): 54-61.

Bank, M., Duffy, F., Leyendecker, V., & Silva, M. (2021).

The Lobby Network: Big Tech's Web of Influence in the EU, Brussels and Cologne: Corporate Europe Observatory and LobbyControl, https://corporateeurope.org/en/2021/08/lobby-network-big-techs-web-influence-eu (accessed 27 August 2021).

Barcelona Declaration (2017). Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe, IIIA CSIC, https://www.iiia.csic.es/barcelonadeclaration/ (accessed 27 August 2021).

Bathaee, Y. (2018). The Artificial Intelligence Black Box and The Failure of Intent and Causation, Harvard Journal of Law & Technology, 31(2): 889-920.

Belmonte, E. (2019). La aplicación del bono social del Gobierno niega la ayuda a personas que tienen derecho a ella, Civio, https://civio.es/tu-derecho-a-saber/2019/05/16/la-aplicacion-del-bono-social-del-gobierno-niega-la-ayuda-a-personas-que-tienen-derecho-a-ella/ (accessed 27 August 2021).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? En Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623), https://dl.acm.org/doi/10.1145/3442188.3445922 (accessed 27 August 2021).

Bietti, E. (2020). From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 210-219), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3513182 (accessed 27 August 2021).

Bigas, E.; Duran, N.; Fuster, E.; Parra, C.; Fernández, T.; Marco, D.; Santanach, D.; Balbuena, E. & Sabater, A. (2021). Anàlisi de l'especialització en intel·ligència artificial. Col·lecció "Monitoratge de la RIS3CAT", número 13, Direcció General de Promoció Econòmica, Competència i Regulació, Generalitat de Catalunya.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorell, T., Wallis, M., Whitby, B., and Winfield, A. (2017). Principles of Robotics: Regulating Robots in the Real World. Connection Science, 29(2): 124-29.

Bryman, A. (2016). Social Research Methods. Oxford: Oxford University Press.

Buolamwini, J., and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, 81, 77–91, https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf (accessed 27 August 2021).

Buolawmini, J. (2018). Project Overview: Gender Shades. MIT Media Lab. https://www.media.mit.edu/projects/gender-shades/overview/ (accessed 27 August 2021).

Butterworth, M. (2018). The ICO and artificial intelligence: The role of fairness in the GDPR framework. Computer Law & Security Review, 34(2): 257-268.

Calo, R. (2017). Artificial Intelligence Policy: A Primer and Roadmap, University of California Davis Rev, 51, 399, https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Calo.pdf (accessed 27 August

2021).

CATALONIA. AI (2020). L'Estrategia d'Intel·ligència Artifical de Catalunya, Generalitat de Catalunya, Departament de Vicepresidència i de Polítiques Digitals i Territori, https://politiquesdigitals.gencat.cat/ca/tic/catalonia-ai (accessed 27 August 2021).

Cellan-Jones, R. (2014). Stephen Hawking warns Artificial Intelligence could end mankind, BBC Technology, https://www.bbc.com/news/technology-30290540 (accessed 27 August 2021).

Comissionat d'Innovació Digital, Administració Electrònica i Bon Govern (2021). Mesura de Govern de l'estrategia municipal d'algoritmes i dades per a l'impuls ètic de la intel·ligència artificial. Barcelona: Ajuntament de Barcelona, https://bcnroc.ajuntament.barcelona.cat/jspui/handle/11703/121795 (accessed 27 August 2021).

Copeland (2021). Artificial Intelligence, Britannica, https://www.britannica.com/technology/artificial-intelligence (accessed 27 August 2021).

Cortés, U., Cortés, A., Barrué, C., Sánchez, A., Moya-Sánchez, E.U., & Garcia-Gasulla, D. (2021). To Be fAIr or Not to Be: Using AI for the Good of Citizens, IEEE Technology and Society Magazine, 40(1): 55-70.

COTEC (2020). III Encuesta COTEC sobre percepción social de la innovación en la sociedad española. Informe del estudio cuantitativo, https://cotec.es/proyecto/iii-encuesta-cotec-sobre-percepcion-social-de-la-innovacion/ (accessed 27 August 2021).

Council of Europe (2021) AI Initiatives. Data Visualisation of AI Initiatives. Strasbourg: Council of Europe Portal, https://www.coe.int/en/web/artificial-intelligence/national-initiatives (accessed 27 August 2021).

Crawford, K. (2021). Atlas of AI – Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven and London: Yale University Press.

Danaher, J., & Robbins, S. (2020). Should AI Be Explainable? [Archivo de audio] Philosophical Disquisitions, https://philosophicaldisquisitions.blogspot.com/2020/07/77-should-ai-be-explainable.html (accessed 27 August 2021).

Daniels, J., Nkonde, M., & Mir, D. (2019). Advancing Racial Literacy in Tech, Data & Society Blog, https://datasociety.net/output/advancing-racial-literacy-in-tech/ (accessed 27 August 2021).

Dastin, J. (2018). Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women, Reuters, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (accessed 27 August 2021).

Dehghani, M., Forbus, K., Tomai, E., & Klenk, M. (2011). An Integrated Reasoning Approach to Moral Decision Making. En M. Anderson and S. L. Anderson (Eds.), Machine Ethics, 422–41. Cambridge: Cambridge University Press.

Dinh, T. N., & Thai, M. T. (2018). AI and Blockchain: A Disruptive Integration, Computer, 51(9): 48-53.

Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018). A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. En 2018 Workshop on Fairness, Accountability and Transparency in Machine Learning during

ICMI, Stockholm, Sweden, https://arxiv.org/abs/1807.00553 (accessed 27 August 2021).

Dutton, T. (25 de julio de 2018) An Overview of National AI Strategies, Medium, https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd (accessed 27 August 2021).

Dwivedi, Y. K., Hughes, L., Ismagilova, E., et al. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy, International Journal of Information Management, 57, 101994.

ECNL (2021) Position statement on the EU AI Act. ECNL - European Center for Not-For-Profit Law, https://ecnl.org/news/ecnl-position-statement-eu-ai-act (accessed 27 August 2021).

ECNL (2021). Evaluating the Risk of AI Systems to Human Rights – ECNL proposal, The Hague, Netherlands: European Center for Non-for-Profit Law Stichting, https://ecnl.org/news/evaluating-risk-ai-systems-human-rights-tier-based-approach (accessed 27 August 2021).

EDPB and EDPS (2021). Call for ban on use of AI for automated recognition of human features in publicly accessible spaces, and some other uses of AI that can lead to unfair discrimination. Press release (21 de junio de 2021). Brussels: European Data Protection Board and European Data Protection Supervisor, https://edpb.europa.eu/news/news/2021/edpb-edps-call-ban-use-ai-automated-recognition-human-features-publicly-accessible_en (accessed 27 August 2021).

ENIA (2020) Estrategia Nacional de Inteligencia Artificial (1.0), Ministerio de Asuntos Económicos y Transformación Digital, https://portal.mineco.gob.es/RecursosNoticia/mineco/prensa/noticias/2020/201202_np_eniav.pdf (accessed 27 August 2021).

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. New York: St. Martin's Press.

European Commission (2019). Ethics guidelines for trustworthy AI, High-Level Expert Group on Artificial Intelligence. Brussels: European Commission, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed 27 August 2021).

European Commission (2020). White Paper on Artificial Intelligence: A European approach to excellence and trust (COM(2020) 65). Brussels: European Commission, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (accessed 27 August 2021).

European Commission (2021). Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN (accessed 27 August 2021).

Floridi, L. (2020). What the Near Future of Artificial Intelligence Could Be, Philosophy & Technology, 32: 1-15.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R.,

Chazerand, P., Dignum, V. & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4): 689-707.

Frey, C. B., & Osborne, M. (2015). Technology at Work - The Future of Innovation and Employment, Oxford Martin School, University of Oxford, https://www.oxfordmartin.ox.ac.uk/downloads/reports/Citi_GPS_Technology_Work.pdf (accessed 27 August 2021).

Freudenburg, W. R. (1988). Perceived risk, real risk: Social science and the art of probabilistic risk assessment, Science, 242(4875): 44-49.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114, 254-280.

Friedman, G. (2014). Workers without employers: Shadow corporations and the rise of the gig economy, Review of Keynesian Economics, Edward Elgar Publishing, 2(2): 171-188.

Garnelo, M., & Shanahan, M. (2019). Reconciling deep learning with symbolic artificial intelligence: representing objects and relations, Current Opinion in Behavioral Sciences, 29: 17-23.

Goffi, E. R. (2021). El 'blanqueo ético' de la tecnología del futuro. La Vanguardia, https://www.lavanguardia.com/internacional/vanguardia-dossier/revista/20210107/6131993/blanqueo-etico-tecnologia-futuro.html (accessed 27 August 2021).

Gordon, J.S. (2020a). Building Moral Machines: Ethical Pitfalls and Challenges, Science and Engineering Ethics, 26, 141-57.

Gordon, J.S. (2020b). What Do We Owe to Intelligent Robots? AI & Society, 35, 209-23.

Gordon, J.S. (2020c). Artificial Moral and Legal Personhood, AI & Society, 36, 457-471.

Grandy, G., & Sliwa, M. (2017). Contemplative leadership: The possibilities for the ethics of leadership theory and practice, Journal of Business Ethics, 143(3): 423-440.

Guarini, M. (2006). Particularism and the Classification and Reclassification of Moral Cases, IEEE Intelligent Systems, 21(4): 22-28.

Gunkel, D. (2018). Robot Rights. Cambridge, MA: MIT Press.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines, Minds and Machines, 30(1): 99-120.

Haldane, A.G. (2015). How low can you go. Speech delivered at the Portadown Chamber of Commerce on 18 September 2015 by Andrew G. Haldane, Chief Economist, Bank of England, https://cryptochainuni.com/wp-content/uploads/Bank-of-England-Speech-given-by-Andrew-G-Haldane-Chief-Economist.pdf (accessed 27 August 2021).

Haner, J., & Garcia, D. (2019). The artificial intelligence arms race: trends and world leaders in autonomous weapons development, Global Policy, 10(3): 331-337.

Henin, C., & Le Métayer, D. (2021). A framework to contest and justify algorithmic decisions, AI and

Ethics, 1-14.

Hewson, C. (2010). Internet-mediated research and its potential role in facilitating mixed methods research. In S. N. Hesse-Biber & P. Leavy (Eds.), Handbook of Emergent Methods (pp. 543–570). New York: Guilford.

Himma, K., & Tavani, H. (2008). The Handbook of Information and Computer Ethics. Hoboken, NJ: Wiley.

Hitzler, P., Bianchi, F., Ebrahimi, M., & Sarker, M. K. (2020). Neural-symbolic integration and the Semantic Web, Semantic Web, 11(1): 3-11.

Hueso, L. C. (2019). Ética en el diseño para el desarrollo de una inteligencia artificial, robótica y big data confiables y su utilidad desde el Derecho, Revista catalana de dret públic, 58, 29-48.

Huws, U. (2014). Labor in the Global Digital Economy: The Cybertariat Comes of Age. New York: New York University Press.

IEEE (2019). Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf (accessed 27 August 2021).

Jing, S., & Doorn, N. (2020). Engineers' Moral Responsibility: A Confucian Perspective, Science and Engineering Ethics, 26(1): 233-253.

Jobin, A. (2020). Ethics guidelines galore for AI – so now what? En ETH Zurich, https://ethz.ch/en/news-and-events/eth-news/news/2020/01/ethics-guidelines-galore-for-ai.html (accessed 27 August 2021).

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines, Nature Machine Intelligence, 1(9): 389-399.

Johnson, D., & Nissenbaum, H. (1995). Computing, Ethics, and Social Values. Englewood Cliffs NJ: Prentice Hall.

Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., ... & Shestakofsky, B. (2021). Toward a Sociology of Artificial Intelligence: A Call for Research on Inequalities and Structural Change, Socius: Sociological Research for a Dynamic World, 7: 1-11.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence, Business Horizons, 62(1): 15-25.

Kraemer, F., Van Overveld, K., & Peterson, M. (2011). Is There an Ethics of Algorithms? Ethics and Information Technology, 13, 251-60.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2020). How We Analyzed the COMPAS Recidivism Algorithm. ProPublica, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed 27 August 2021).

Lee, K.F. (2018). AI Superpowers: China, Silicon Valley, and the New World Order. Boston: Houghton Mifflin Harcourt.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., &

Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes, Philosophy & Technology, 31(4): 611-627.

Lin, P., Abney, K. & Bekey, G. A. (Eds). (2014). Robot Ethics: The Ethical and Social Implications of Robotics. Intelligent Robotics and Autonomous Agents. Cambridge, MA and London: MIT Press.

Lin, P., Abney, K., & Jenkins, R. (Eds.) (2017). Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence. New York: Oxford University Press.

Lozano, I. A., Molina, J. M., & Gijón, C. (2021). Perception of Artificial Intelligence in Spain. Telematics and Informatics, 63, 101672.

Lu, M. (2020). 50 Cognitive Biases in the Modern World. Visual Capitalist, https://www.visualcapitalist.com/50-cognitive-biases-in-the-modern-world/ (accessed 27 August 2021).

Ludwig, S. (2015). Credit Scores in America Perpetuate Racial Injustice: Here's How. The Guardian, https://www.theguardian.com/commentisfree/2015/oct/13/your-credit-score-is-racist-heres-why (accessed 27 August 2021).

Lum, K. (2021). What is an "algorithm"? It depends whom you ask. MIT Technology Review, https://www.technologyreview.com/2021/02/26/1020007/what-is-an-algorithm/ (accessed 27 August 2021).

Martínez, C., Skeet, A. G., & Sasia, P. M. (2021). Managing organizational ethics: How ethics becomes pervasive within organizations, Business Horizons, 64(1), 83-92.

McCarthy, J. (2007). What is Artificial Intelligence? Stanford University, http://jmc.stanford.edu/articles/whatisai/whatisai.pdf (accessed 27 August 2021).

McKenna, M. (2019). Machines and Trust: How to Mitigate AI Bias. Toptal Engineering Blog, https://www.toptal.com/artificial-intelligence/mitigating-ai-bias (accessed 27 August 2021).

Metadata (2021) Catalunya, capdavantera en intel·ligència artificial a Europa. Metadata, https://www.metadata.cat/noticia/918/catalunya-intelligencia-artificial-europa-regions-capdavanteres (accessed 27 August 2021).

Metzinger, T. (2019). Ethics Washing Made in Europe. Der Tagesspiegel, https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html (accessed 27 August 2021).

Mikalef, P., Framnes, V. A., Danielsen, F., Krogstie, J., & Olsen, D. (2017). Big Data Analytics Capability: Antecedents and Business Value. In Proceeding of the Pacific Asia Conference on Information Systems (PACIS), https://aisel.aisnet.org/pacis2017/136 (accessed 27 August 2021).

Misuraca, G., & van Noordt, C. (2020). AI Watch – Artificial Intelligence in public services –Overview of the use and impact of AI in public services in the EU. Seville: European Commission's Joint Research Centre, Digital Economy Unit, https://publications.jrc.ec.europa.eu/repository/handle/JRC120399 (accessed 27 August 2021).

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate, Big Data & Society, 3(2): 1-21.

Moss, E., Watkins, E. A., Singh, R., Elish, M. C., & Metcalf,

J. (2021). Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. Data & Society, https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf (accessed 27 August 2021).

Müller-Eiselt, R., & Hustedt, C. (2020) Algo.Rules - From principles to practice: How can we make AI ethics measurable? Bertelsmann Stiftung, https://www.bertelsmann-stiftung.de/en/our-projects/ethics-of-algorithms/project-news/from-principles-to-practice-how-can-we-make-ai-ethics-measurable (accessed 27 August 2021).

Neri, H., & Cozman, F. (2020). The role of experts in the public perception of risk of artificial intelligence, AI & SOCIETY, 35(3): 663-673.

Neudert, L. M., Knuutila, A., & Howard, P. (2020). Global Attitudes towards AI, Machine Learning & Automated Decision Making, Oxford Commission of AI & Good Governance. Oxford: Oxford Internet Institute / University of Oxford, https://oxcaigg.oii.ox.ac.uk/wp-content/uploads/sites/124/2020/10/GlobalAttitudesTowardsAIMachineLearning2020.pdf (accessed 27 August 2021).

Nyholm, S. (2020). Humans and Robots: Ethics, Agency, and Anthropomorphism. London: Rowman and Littlefield.

Nyholm, S., & Frank, L. (2019). It Loves Me, It Loves Me Not: Is It Morally Problematic to Design Sex Robots That Appear to Love Their Owners? Techne: Research in Philosophy and Technology, 23(3), 402-24.

O'Brien, C. (2020) Facebook civil rights audit urges 'mandatory' algorithmic bias detection. VentureBeat, https://venturebeat.com/2020/07/08/facebook-civil-rights-audit-urges-mandatory-algorithmic-bias-detection/ (accessed 27 August 2021).

O'Neil, C. (2016). Weapons of Math Destruction. London: Allen Lane.

Observatorio Nacional de Tecnología y la Sociedad (2021). Indicadores de uso de Inteligencia Artificial en las empresas españolas. Madrid: Ministerio de Asuntos Económicos y Transformación Digital, Secretaria General Técnica, https://www.ontsi.red.es/es/dossier-de-indicadores-pdf/indicadores-uso-inteligenciaartificialempresas-espanolas (accessed 27 August 2021).

OECD (2019). Recommendation of the Council on Artificial Intelligence, Committee on Digital Economy Policy, OECD Legal Instruments, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#committees (accessed 27 August 2021).

OECD (2021). OECD AI Policy Observatory, https://oecd.ai. (accessed 27 August 2021).

Ortega Klein, A. (2020). Geopolítica de la ética en Inteligencia artificial. Documento de trabajo 1/2020. Madrid: Real Instituto Elcano, http://www.realinstitutoelcano.org/wps/portal/rielcano_es/contenido?WCM_GLOBAL_CONTEXT=/elcano/elcano_es/zonas_es/dt1-2020-ortega-geopolitica-de-la-etica-en-inteligencia-artificial (accessed 27 August 2021).

Picard, R. (1997). Affective Computing. Cambridge, MA and London: MIT Press.

Pichai, S. (2018). AI at Google: our principles, https://blog.google/technology/ai/ai-principles/ (accessed 27 August 2021).

Reinhold, F., & Müller, A. (2021). AlgorithmWatch's response to the European Commission's proposed regulation on Artificial Intelligence – A major step with major gaps. AlgorithmWatch. https://algorithmwatch.org/en/response-to-eu-ai-regulation-proposal-2021/ (accessed 27 August 2021).

Resseguier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data & Society, 7(2), 2053951720942541.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144).

Robbins, S. (2019). A Misdirected Principle with a Catch: Explicability for AI, Minds and Machines, 29(4), 495–514.

Rothschild, J. (2016). The Logic of a Co-operative Economy and Democracy 2.0: Recovering the Possibilities for Autonomy, Creativity, Solidarity, and Common Purpose, The Sociological Quarterly, 57(1), 7-35.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence, 1(5), 206-215.

Russel, S. & Norvig (Eds) (2003). Artificial Intelligence: A Modern Approach. Hoboken: Pearson Education

Salas, J. (2019). El temor a la inteligencia artificial surge del recelo hacia los intereses económicos. El País, https://elpais.com/elpais/2019/11/14/ciencia/1573728249_279206.html (accessed 27 August 2021).

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021). AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection, IEEE Transactions on Technology and Society, 2(1): 31-42.

Schneider, S., & Leyer, M. (2019). Me or information technology? Adoption of artificial intelligence in the delegation of personal strategic decisions, Managerial and Decision Economics, 40(3): 223-231.

Shen, B. (2018). Cómo mitigar los sesgos injustos en la inteligencia artificial. OpenGlobalRights, https://www.openglobalrights.org/mitigating-unfair-bias-in-artificial-intelligence/?lang=Spanish (accessed 27 August 2021).

Smith, R. E. (2019). Rage Inside the Machine: The Prejudice of Algorithms, and How to Stop the Internet Making Bigots of Us All. London: Bloomsbury Academic.

Spencer, D. A. (2018). Fear and hope in an age of mass automation: debating the future of work, New Technology, Work and Employment, 33(1): 1-12

.

Springer, A., Garcia-Gathright, J. & Cramer, H. (2018). Assessing and Addressing Algorithmic Bias – But Before We Get There. In 2018 AAAI Spring Symposium Series, 450–54. https://www.aaai.org/ocs/index.php/SSS/SSS18/paper/viewPaper/17542 (accessed 27 August 2021).

Steels, L., & López de Mantaras, R. (2018). The Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe, AI Communications, 31(6): 485-494.

Sweeney, L. (2013). Discrimination in Online Ad Delivery, Acmqueue, 11(3): 1-19.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good, Science, 361(6404): 751-752.

Turing, A. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem, Proceedings of the London Mathematical Society, s2-42(1): 230-265.

Turing, A. (1950). Computing Machinery and Intelligence, Mind, LIX(236): 433-460.

UNESCO (2019). Preliminary report on the first draft of the Recommendation on the Ethics of Artificial Intelligence, https://en.unesco.org/artificial-intelligence/ethics(accessed 27 August 2021).

Van Roy, V., Rossetti, F., Perset, K., & Galindo-Romero, L. (2021). AI Watch-National strategies on Artificial Intelligence: A European perspective. Seville: European Commission's Joint Research Centre, Digital Economy Unit, https://publications.jrc.ec.europa.eu/repository/handle/JRC122684 (accessed 27 August 2021).

Veale, M., & Binns, R. (2017). Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data. Big Data & Society, 4(2): 1-17.

Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach, Computer Law Review International, 22(4): 97-112.

Véliz, C. (2020). Privacy is Power. Why and How You Should Take Back Control of Your Data. Ealing, London: Transworld Publishers.

VVAA (2020). Reclaim Your Face, https://reclaimyourface.eu/ (accessed 27 August 2021).

Wachter, S., Mittelstadt, B. and Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology, 31(2): 841–87.

Wagner, B. (2018). Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping? Being Profiled (pp. 84-89). Amsterdam: Amsterdam University Press.

Wallach, W., & Allen, C. (2010). Moral Machines: Teaching Robots Right from Wrong. Oxford: Oxford University Press.

Wallach, W., Franklin, S., & Allen, C. (2010). A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents, Topics in Cognitive Science, 2(3): 454-485.

Wang, Y., & Kosinski, M. (2018). Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images, Journal of Personality and Social Psychology, 114(2): 246-257.

Wareham, C. S. (2020). Artificial intelligence and African conceptions of personhood. Ethics and Information Technology, 23, 127-136.

Williams, J., & Chowdhury, R. (2021). Introducing our Responsible Machine Learning Initiative, https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative (accessed 27 August 2021).

Williams, O. (2019). How Big Tech funds the debate on AI ethics. New Statesman, https://www.newstatesman.com/science-tech/technology/2019/06/how-big-tech-funds-debate-ai-ethics (accessed 27 August 2021).

Wirtz, B. W., & Müller, W. M. (2019). An integrated artificial intelligence framework for public management, Public Management Review, 21(7): 1076-1100.

World Economic Forum (2021). Responsible Use of Technology: The Microsoft Case Study. En colaboración con Markkula Center for Applied Ethics. http://www3.weforum.org/docs/WEF_Responsible_Use_of_Technology_2021.pdf (accessed 27 August 2021).

World Health Organization (2021). WHO consultation towards the development of guidance on ethics and governance of artificial intelligence for health. Meeting report, https://www.who.int/publications/i/item/who-consultation-towards-the-development-of-guidance-on-ethics-and-governance-of-artificial-intelligence-for-health (accessed 27 August 2021).

Annex 1. Semi-structured interview script

Introduction

1. To situate yourself, what discipline or field of artificial intelligence (AI) do you work in or have an interest in?

2. What is the first word or adjective that comes to mind when we say AI?

Ethical and social domain

1. Looking to the present but also to the future, do you think ethical AI should be seen as a constraint on AI actions, as a sub-goal or as the main goal?

2. To what extent do you agree with the following statement: People are not computer processes or programmes, but unique individuals with empathy, self-determination, unpredictability, intuition and creativity and therefore have a higher status than machines? (Answer on a Likert scale 1-5, where 1 strongly disagrees and 5 strongly agrees).

3. In your opinion, do you think AI will weaken or discourage some important human habits, skills or virtues that are fundamental to human excellence (moral, political or intellectual)?

4. Conversely, do you think AI will strengthen some important human habits, skills or virtues that are fundamental to human excellence (moral, political or intellectual)?

5. In your opinion, do you think we need to consider the ethical perspectives of AI recipients and communities other than our own, including those who are culturally or physically far away from us?

6. How do you think younger generations may be affected by the widespread use of AI systems?

7. To what extent do you agree with the following statement: In view of the widespread implementation of AI systems, a wide range of actions are being considered

resources/opportunities to increase potential benefits and minimise risks for younger generations? (Likert scale response 1-5, where 1 strongly disagree and 5 strongly agree).

8. In your opinion, what are the repercussions globally and also by specific regions in Europe, when there are countries that invest heavily in AI (e.g. China) and do not prohibit or restrict the technological development of AI on the same terms as others or as we do from Europe?

Legal domain

1. In your opinion, who is or should be responsible for setting and enforcing ethical standards for AI systems?

2. Which type of AI regulation do you currently consider more appropriate, restrictive (the regulate-and-forget type) or adaptive (the iterative with technological change type)?

3. To what extent do you agree with the following statement: the responsibility for an AI decision, action and process should always be taken by a natural or legal person? (Answer on a Likert scale 1-5, where 1 strongly disagrees and 5 strongly agrees).

4. To what extent do you agree with the following statement: sustainable processing of personal data should ensure accountability in the short, medium and long term. Context: commercial and government data that is accumulated over time allows for an incredibly detailed portrait of an individual's life? (Answer on a Likert scale 1-5, with 1 being strongly disagree and 5 being strongly agree).

5. In your opinion, how can we ensure that the algorithms used in AI systems are fair, especially when they are privately owned by corporations and not accessible to public scrutiny?

6. To what extent do you agree with the following statement: the use of biometric surveillance technologies (e.g., remote facial recognition) used indiscriminately or arbitrarily in publicly accessible spaces represents a violation of people's fundamental rights and freedoms? (Likert scale 1-5, where 1 is strongly disagree and 5 is strongly agree).

7. If you had to choose between more precise and secure algorithms or transparent algorithms, which would you choose? Highlight if there is a trade-off or consequence.

8. To what extent do you agree with the following statement: AI predictive models for predicting where and by whom certain types of crime are likely to be committed over-represent poor, working class, racialised or migrant communities with a higher presumptive likelihood of future criminality, and this represents a violation of people's fundamental rights and freedoms (Likert scale answer 1-5).

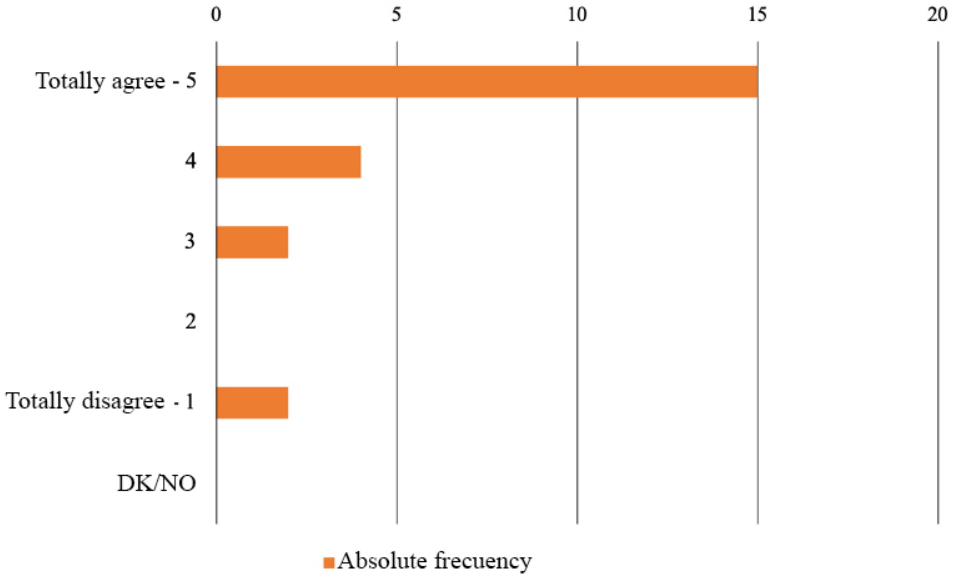
Looking to the future

1. Looking at a long-term (25-year) future trajectory, perhaps in a general AI context, what do you think will be the main ethical and social challenges that will cause, for example, governments to oversee, shut down and/or nationalise AI systems?

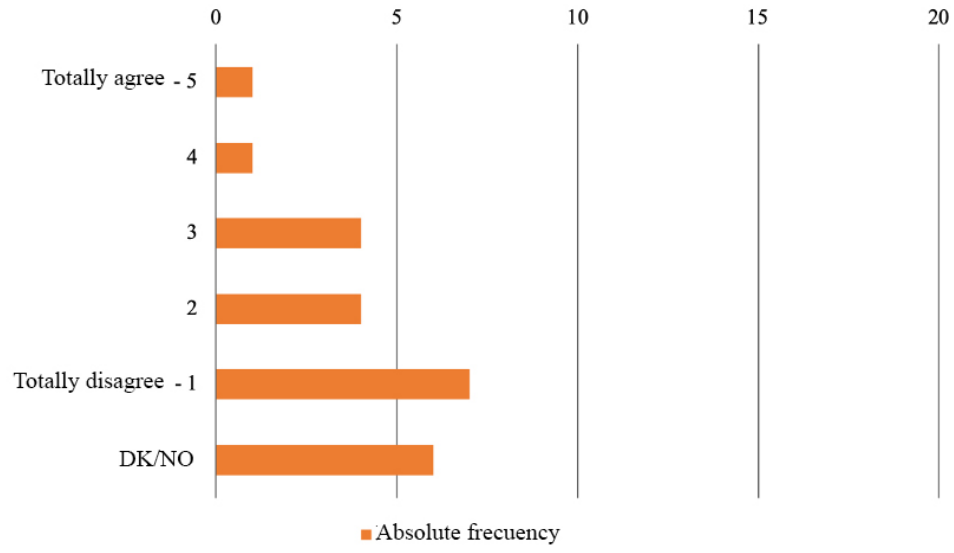
2. In your field, do you think that the advantages or opportunities for AI development will outweigh the major drawbacks or risks in ethical and social terms?

Annex 2. Graphical results of the closed-ended interview questions

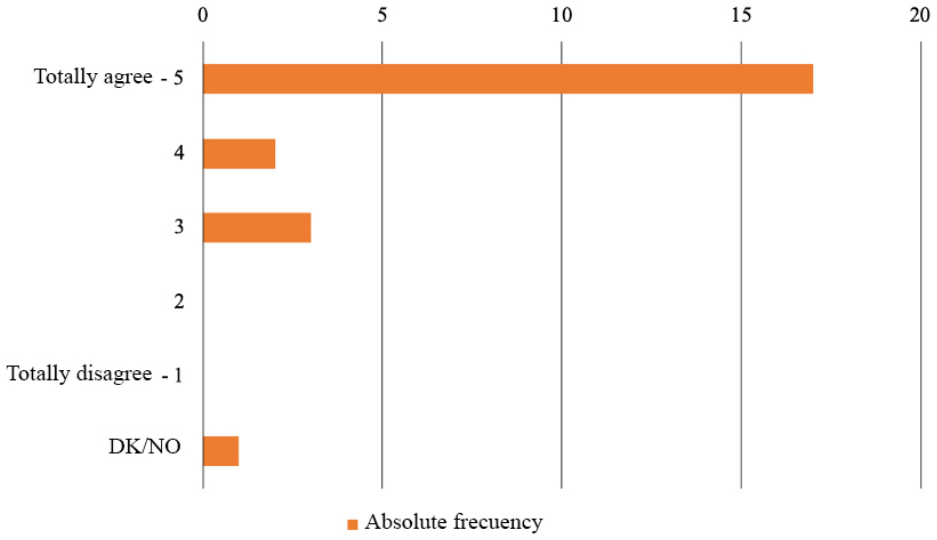
- To what extent do you agree with the following statement: people are not computer processes or programmes, but unique with empathy, self-determination, unpredictability, intuition and creativity and therefore have a higher status than machines? (Answer on a Likert scale of 1-5).



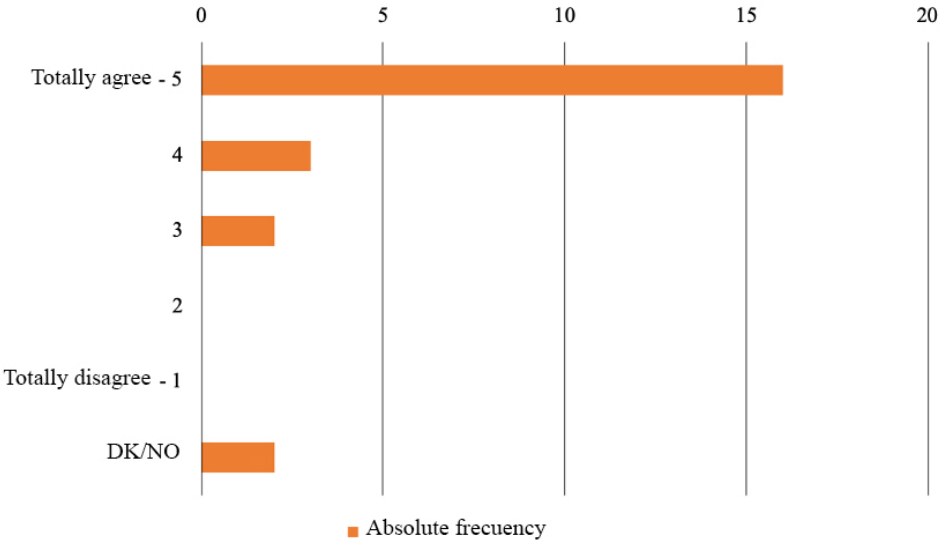
- ¿Hasta qué punto estáis de acuerdo con la siguiente afirmación: Ante la implantación generalizada de sistemas de IA se están considerando una amplia gama de acciones / recursos / oportunidades para aumentar los beneficios potenciales y minimizar los riesgos para las generaciones más jóvenes? (Respuesta en una escala Likert de 1-5).



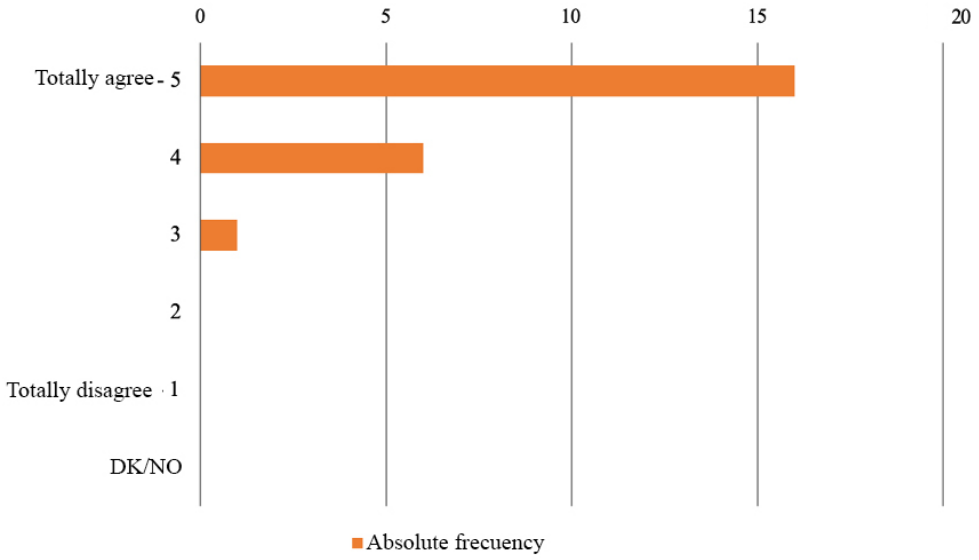
- ¿Hasta qué punto estáis de acuerdo con la siguiente afirmación: La responsabilidad de una decisión, acción y proceso de IA tiene que ser asumida siempre por una persona física o jurídica? (Respuesta en una escala de Likert 1-5).



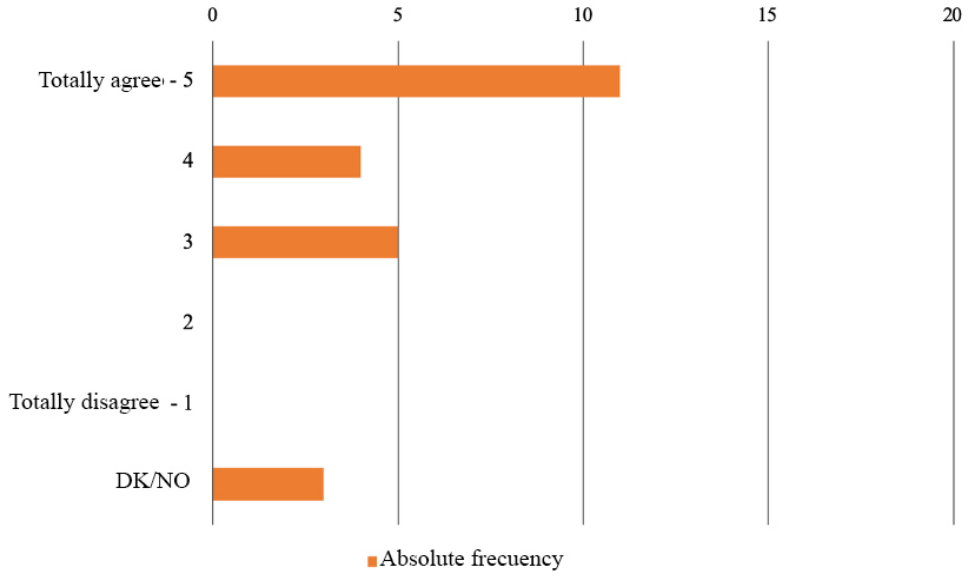
- ¿Hasta qué punto estáis de acuerdo con la siguiente afirmación: El procesamiento sostenible de datos personales tiene que garantizar una rendición de cuentas a corto, medio y largo plazo? (Respuesta en una escala de Likert 1-5).



- ¿Hasta qué punto estáis de acuerdo con la siguiente afirmación: el uso de tecnologías de vigilancia biométrica (por ejemplo, el reconocimiento facial remoto) utilizadas de manera indiscriminada o arbitraria en espacios accesibles públicamente representa una violación de los derechos y las libertades fundamentales de las personas? (Respuesta en una escala de Likert 1-5).



- ¿Hasta qué punto estáis de acuerdo con la siguiente afirmación: Los modelos predictivos de la IA para predecir donde y por quién es probable que se cometan ciertos tipos de delitos sobrerepresenten a comunidades pobres, de clase trabajadora, racializadas y migradas con una mayor probabilidad de presuntiva criminalidad futura y esto representa una violación de los derechos y las libertades fundamentales de las personas? (Respuesta en una escala de Likert 1-5).



Annex 3. People interviewed

Ariel Guersenzvaig
Elisava Escuela de Diseño e Ingeniería de Barcelona

Artur Serra
Fundació i2CAT

Carina Lopes
Digital Future Society

Carme Torras
Institut de Robòtica i Informàtica Industrial, Consell Superior d'Investigacions Científiques, Universitat Politècnica de Catalunya

David Pereira
Everis

Elisabet Golobardes
La Salle, Universitat Ramon Llull

Fernando Vilariño
Centre de Visió per Computador, Universitat Autònoma de Barcelona

Itziar de Lecuona
Observatori de Bioètica i Dret, Universitat de Barcelona

Joan Manuel del Pozo
Professor emèrit i Síndic de la Universitat de Girona

Joan Mas
Centre of Innovation for Data tech and Artificial Intelligence (CIDAI)

Karina Gibert
Intelligent Data Science and Artificial Intelligence Research Center, Universitat Politècnica de Catalunya

Karma Peiró
Journalist and Co-director of Fundació Visualització per la Transparència

Liliana Arroyo
ESADE

Marc Pérez-Batlle
Institut Municipal d'Informàtica, Ajuntament de Barcelona

Miquel Domènech
Departament de Psicologia Social, Universitat Autònoma de Barcelona

Montse Guàrdia
General Manager of Alastria Blockchain Ecosystem

Joaquim Meléndez
eXIT Research Group, Control Engineering and Intelligent Systems, Universitat de Girona

Ramon López de Mántaras
Institut d'Investigació en Intel·ligència Artificial, Consell d'Investigacions Científiques, Universitat Autònoma de Barcelona

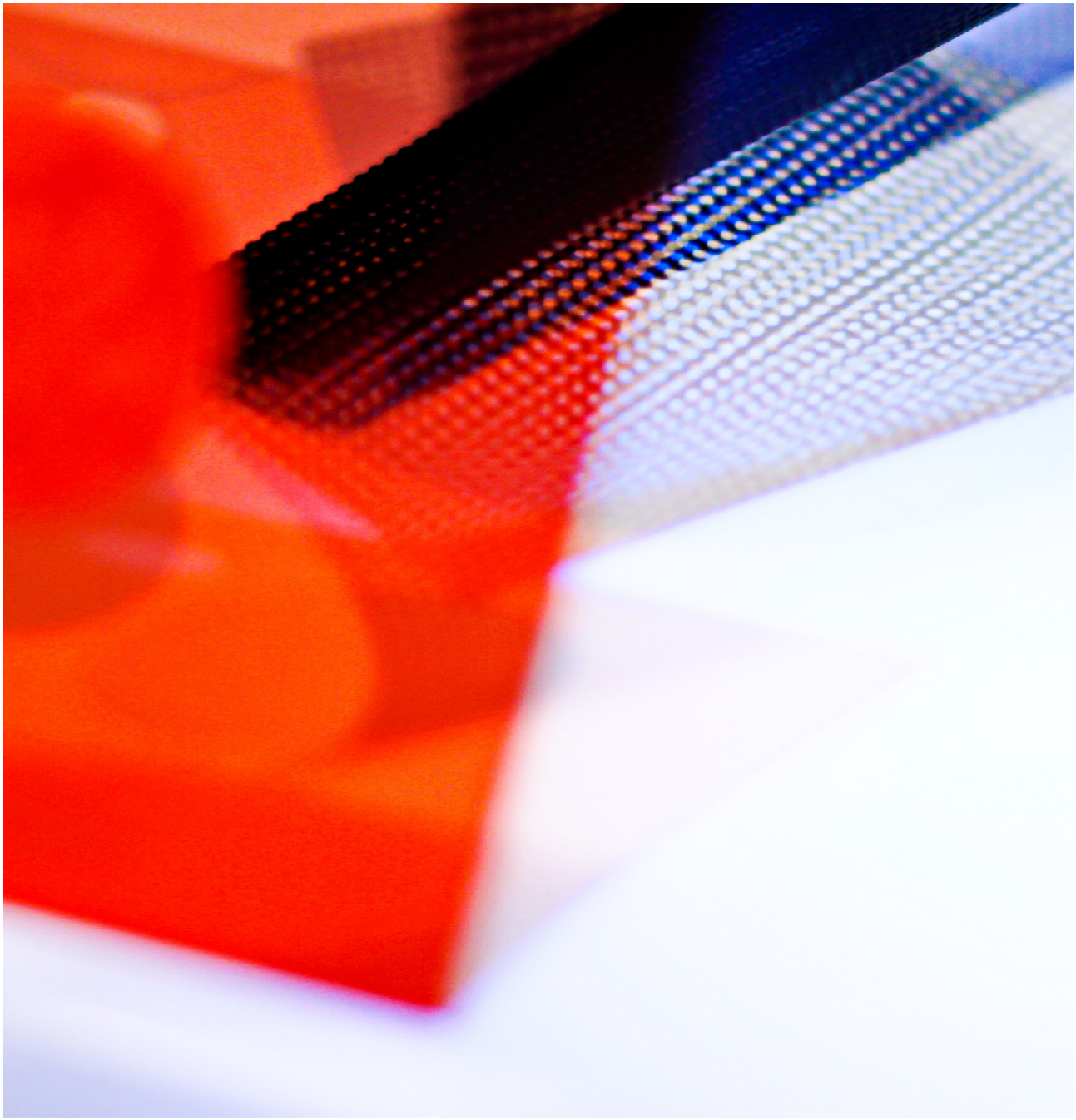
Ramon Trias
President and Director of AIS Group

Ricardo Baeza-Yates
Web Science & Social Computing Group, Universitat Pompeu Fabra

Ulises Cortés
Barcelona Supercomputing Center, Centro Nacional de Supercomputación, Universitat Politècnica de Catalunya

Xavier Marcet
Lider of Lead to Change

Xavier Trabado
m4Social



OBSERVATORI D'ÈTICA EN INTEL·LIGÈNCIA ARTIFICIAL DE CATALUNYA

www.oeiac.cat suport.oeiac@udg.edu @OEIAC_UdG

Generalitat de Catalunya
Departament de la Vicepresidència
i de Polítiques Digitals i Territori

i2cat

Universitat de Girona